Méthodes particulaires et vraisemblances pour l'inférence de modèles d'évolution avec dépendance au contexte

Alexis Huet

N° d'ordre : 126-2014

Université de Lyon
Université Claude Bernard Lyon 1
Institut Camille Jordan
École doctorale InfoMaths

THÈSE

Spécialité mathématiques

en vue d'obtenir le grade de docteur,
présentée et soutenue publiquement par

Alexis HUET

le 27 juin 2014

Méthodes particulaires et vraisemblances pour l'inférence de modèles d'évolution avec dépendance au contexte

---

Thèse encadrée par

M. Jean BÉRARD
Mme Anne-Laure FOUGÈRES

soutenue après avis de

Mme Catherine MATIAS
M. Eric MOULINES

devant la commission d'examen formée de

M. Jean BÉRARD	*directeur de thèse*
M. Michael BLUM	*examinateur*
Mme Anne-Laure FOUGÈRES	*directrice de thèse*
M. Nicolas LARTILLOT	*examinateur*
Mme Catherine MATIAS	*rapporteuse*
M. Eric MOULINES	*rapporteur*
M. Didier PIAU	*examinateur*

Résumé

Cette thèse est consacrée à l'inférence de modèles stochastiques d'évolution de l'ADN avec dépendance au contexte, l'étude portant spécifiquement sur la classe de modèles stochastiques RN95+YpR. Cette classe de modèles repose sur un renforcement des taux d'occurrence de certaines substitutions en fonction du contexte local, ce qui introduit des phénomènes de dépendance dans l'évolution des différents sites de la séquence d'ADN. Du fait de cette dépendance, le calcul direct de la vraisemblance des séquences observées met en jeu des matrices de dimensions importantes, et est en général impraticable.

Au moyen d'encodages spécifiques à la classe RN95+YpR, nous mettons en évidence de nouvelles structures de dépendance spatiales pour ces modèles, qui sont associées à l'évolution des séquences d'ADN sur toute leur histoire évolutive. Ceci rend notamment possible l'utilisation de méthodes numériques particulaires, développées dans le cadre des modèles de Markov cachés, afin d'obtenir des approximations consistantes de la vraisemblance recherchée. Un autre type d'approximation de la vraisemblance, basé sur des vraisemblances composites, est également introduit.

Ces méthodes d'approximation de la vraisemblance sont implémentées au moyen d'un code en C++. Elles sont mises en œuvre sur des données simulées afin d'étudier empiriquement certaines de leurs propriétés, et sur des données génomiques, notamment à des fins de comparaison de modèles d'évolution.

Mots clefs : modèles d'évolution avec dépendance au contexte, chaînes de Markov cachées, méthodes particulaires, filtre particulaire auxiliaire, vraisemblances composites.

Abstract

This thesis is devoted to the inference of context-dependent evolutionary models of DNA sequences, and is specifically focused on the RN95+YPR class of stochastic models. This class of models is based on the reinforcement of some substitution rates depending on the local context, which introduces dependence phenomena between sites in the evolution of the DNA sequence. Because of these dependencies, the direct computation of the likelihood of the observed sequences involves high-dimensional matrices, and is usually infeasible.

Through encodings specific to the RN95+YpR class, we highlight new spatial dependence structures for these models, which are related to the evolution of DNA sequences throughout their evolutionary history. This enables the use of particle filter algorithms, developed in the context of hidden Markov models, in order to obtain consistent approximations of the likelihood. Another type of approximation of the likelihood, based on composite likelihoods, is also introduced.

These approximation methods for the likelihood are implemented in a C++ program. They are applied on simulated data to empirically investigate some of their properties, and on genomic data, especially for comparison of evolutionary models.

Keywords : context-dependent evolutionary models, hidden Markov models, particle filter, auxiliary particule filter, composite likelihood methods.

Remerciements

Tout d'abord, je souhaite remercier mes directeurs de thèse Jean Bérard et Anne-Laure Fougères. Merci d'avoir été présents et disponibles durant ces années, de m'avoir conseillé et soutenu, tant au niveau de la recherche que de la rédaction du manuscrit.

Je remercie ensuite mes rapporteurs Catherine Matias et Eric Moulines. Je suis très honoré que vous ayez accepté d'être mes rapporteurs, de l'intérêt porté à mon manuscrit et des remarques formulées pour l'améliorer. J'adresse aussi mes remerciements à Michael Blum, Nicolas Lartillot et Didier Piau, pour avoir accepté d'être mes examinateurs et pour votre présence à ma soutenance. Je souhaite en particulier remercier Didier de m'avoir donné envie de découvrir le monde des probabilités et de m'avoir initié à la recherche, durant ma première année de master.

Je remercie Laurent Guéguen pour sa patience, pour m'avoir aidé à installer Bio++, à utiliser BppML, et pour m'avoir fourni des alignements de séquences biologiques. Merci également à Hédi Soula, pour les cours de programmation auxquels j'ai pu assister.

Je souhaite remercier tous les doctorants de l'ICJ que j'ai côtoyé pendant ces années lyonnaises. Merci d'abord à tous mes collègues du bureau 219, pour la bonne ambiance pendant ces années. Ceux déjà partis ou prêt à partir vers de nouveaux horizons : Alain, Amélie, Chang, Élodie, JB, 林志聪. Ceux de la nouvelle génération : Benjamin, Corentin, Mathias. Merci aux doctorants/docteurs du premier étage, pour les midis passés ensemble au Domus : Anne, Bérénice, Cécilia, Evrad, Ivan, Nadja, Rudy, Simon, Tomás, Xavier.

Merci à Élodie, Irène, Marielle et Nicolas. J'ai eu beaucoup de chance de vous connaître pendant ma première année à Lyon !

Merci à ceux que j'ai rencontré à Grenoble ou à Lyon, je pense en particulier à Brahim, 郭振中, Jia et Kévin.

Merci aux amis de longue date ! Merci Pierre pour ta vision du monde et ton sens du non-attachement. Merci à tout egatop : JE & Madhu, Julien, Paul & Amandine. N'oubliez pas : « vive le FI » ! Merci Fabrice pour tous les moments passés devant l'ordi, ainsi qu'à Bruno & Muriel, Emilie, Marina, Moïra et Sara. Une pensée aux pierres des les collections de pierres.

Merci à ma famille : à tous ceux qui ont pu assister à ma soutenance mais également à ceux qui n'ont pas pu se libérer. Merci 罗毓 pour m'avoir supporté et apporté ton amour pendant toutes ces années. Merci à mon frère Olivier et à ma belle-sœur 恵子, pour tous les origamis que j'ai pu faire pour m'aérer l'esprit. Merci enfin à mes parents, pour votre soutien depuis toujours.

Table des matières

Introduction **13**

1 Modèles d'évolution **19**
1.1 Modèles à sites indépendants 20
1.1.1 Hypothèses additionnelles 21
1.1.2 Description des modèles T92, RN95 et GTR 22
1.2 Le modèle RN95+YpR 23
1.2.1 Définition du modèle RN95+YpR 24
1.2.2 Gestion de la condition aux bords 26
1.2.3 Construction de l'évolution par processus de Poisson 26
1.2.4 Modèles inclus dans le modèle RN95+YpR 27
1.3 Évolution RN95+YpR sur un arbre 29
1.3.1 Arbre phylogénétique 29
1.3.2 Écriture globale du modèle 31
1.4 Pertinence biologique des hypothèses 34
1.4.1 Alignement de séquences 34
1.4.2 Hypothèses d'homogénéité 34
1.4.3 Hypothèses supplémentaires 35
1.5 Modèles d'évolution généraux 36
1.5.1 Description de la dynamique 36
1.5.2 Espace d'états et description d'une évolution 37
1.5.3 Densité vis-à-vis de la mesure μ 38
1.5.4 Propriétés standards des chaînes de Markov en temps continu 38

2 Vraisemblances pour les modèles RN95+YpR **41**

3 Encodages des séquences et modèles RN95+YpR **47**
3.1 Encodages sur l'alphabet et les séquences 48
3.2 Évolutions encodées et ambigües 48
3.2.1 Évolutions encodées 48
3.2.2 Un cas particulier : l'encodage (ρ, η) 49
3.2.3 Évolutions ambigües 49
3.2.4 Liens avec les évolutions encodées 50
3.3 Problèmes de conditionnement 52
3.4 Précision sur la gestion de la condition aux bords 54
3.5 Conséquences pour le calcul de la vraisemblance 54
3.5.1 Calcul pour les séquences courtes encodées 55

3.5.2 Découpage en produits minimaux 55

4 Vraisemblances composites **57**
4.1 Triplets encodés 60
4.1.1 Construction et propriétés asymptotiques 60
4.1.2 Variance asymptotique du maximum de vraisemblance composite . . 61
4.2 Approximations markoviennes 65

5 Dépendance le long d'une séquence **69**
5.1 Étude d'un modèle d'évolution limite 70
5.1.1 Propriétés du modèle 70
5.1.2 Probabilité d'un nucléotide sachant les observations passées 71
5.1.3 Approximations markoviennes et valeur exacte 72
5.1.4 Inférence à la racine 73
5.2 Loi stationnaire de l'approximation markovienne 76
5.2.1 Définition des quantités recherchées 77
5.2.2 Un modèle extrême 78
5.2.3 Modèles simulés 84

6 Structures markoviennes **85**
6.1 Structure spatiale de champ markovien 86
6.1.1 Structure de champ markovien d'ordre deux 86
6.1.2 Structure de champ markovien d'ordre un 90
6.1.3 Écriture de l'évolution 92
6.1.4 Fonction de survie 95
6.1.5 Structure associée de chaîne de Markov 97
6.2 Structure de chaîne de Markov explicite 99
6.2.1 Évolution en termes d'encodages (ρ, η) 99
6.2.2 Structure spatiale de chaîne de Markov 99
6.2.3 Construction de l'évolution encodée par processus de Poisson 101
6.2.4 Preuves des théorèmes 6.2.2 et 6.2.3 103
6.2.5 Conditionnement par le dinucléotide encodé final 107
6.2.6 Fonction de survie 108
6.3 Structure basée sur le π-encodage 108
6.4 Description des structures sur le modèle complet 111
6.4.1 Description de l'évolution lorsque la racine n'est pas constante 111
6.4.2 Description de l'évolution sur un arbre 112

7 Propriétés du maximum de vraisemblance **115**
7.1 Théorèmes limites pour les chaînes de Markov cachées 116
7.1.1 Définition d'une chaîne de Markov cachée et propriétés 116
7.1.2 Théorèmes de consistance et de normalité asymptotique 117
7.2 Description du noyau de transition 119
7.2.1 Mesure de référence 119
7.2.2 Noyau de transition 122
7.2.3 Écriture sur un arbre 124
7.2.4 Noyau de transition vers les observations 124
7.3 Adaptation du théorème pour notre modèle 125

7.3.1 Condition de Doeblin du noyau Q . . . 125
7.3.2 Vérification des hypothèses du théorème 7.1.9 . . . 128
7.3.3 Hypothèse alternative . . . 129

8 Méthodes de simulation **133**
8.1 Méthodes particulaires générales . . . 134
8.1.1 Algorithme SISR générique . . . 135
8.1.2 Algorithme SISR standard . . . 136
8.1.3 Filtre particulaire auxiliaire . . . 136
8.1.4 Résumé et résultats de convergence et de normalité asymptotique . . . 137
8.1.5 Problèmes de dégénérescence . . . 138
8.2 Méthodes particulaires pour RN95+YpR . . . 141
8.2.1 Calcul des rapports de probabilités . . . 141
8.2.2 Calcul de la vraisemblance . . . 142
8.2.3 Solution aux problèmes de dégénérescence . . . 143
8.2.4 Calcul de la vraisemblance au maximum de vraisemblance . . . 143
8.3 Simulation exacte de la loi stationnaire . . . 145
8.3.1 Dynamique observée à travers les processus de Poisson . . . 146
8.3.2 Verrouillage des sites . . . 146
8.3.3 Algorithme de simulation de la loi stationnaire . . . 147

9 Implémentation **149**
9.1 Programme du point de vue de l'utilisateur . . . 149
9.2 Structure générale du programme . . . 150
9.3 Structure détaillée . . . 152
9.3.1 Calculs matriciels . . . 152
9.3.2 Calculs pour la fonction de survie . . . 153
9.3.3 Description d'une étape de l'avancement d'un site . . . 154
9.3.4 Avancements particulaires pour un site . . . 156

10 Applications **159**
10.1 Comparaison des approximations sur séquences courtes . . . 162
10.1.1 Approximation markovienne à un pas et valeur exacte . . . 162
10.1.2 Approximations particulaires et markoviennes . . . 164
10.2 Comparaison des estimateurs de vraisemblance . . . 166
10.2.1 Approximations particulaires et markoviennes : cas atypiques . . . 166
10.2.2 Approximations particulaires et markoviennes : cas typiques . . . 170
10.2.3 Approximations particulaires avec et sans rééchantillonnage . . . 172
10.3 Fluctuations des approx. par méthodes particulaires . . . 175
10.3.1 Fluctuations des estimations de $p(z_{i+1}(T) \mid z_{1:i}(T))$ . . . 176
10.3.2 Fluctuations des estimations de vraisemblance de la séquence . . . 184
10.3.3 Présence et estimation du biais . . . 186
10.4 Inférence d'un nucléotide de la racine . . . 190
10.4.1 Méthodes d'inférence à la racine . . . 192
10.4.2 Comparaison entre les deux méthodes . . . 194
10.4.3 Influence de la séquence observée - modèle atypique . . . 198
10.4.4 Influence de la séquence observée - modèles typiques . . . 201
10.4.5 Inférence de la séquence complète à la racine . . . 202

10.5 Comparaison des approximations du max. de vraisemblance 203
10.5.1 Comparaison entre les approximations composites et particulaires . . 204
10.5.2 Comparaison entre les approximations composites 205
10.5.3 Exemple d'estimation de la variance associée aux triplets encodés . . 206
10.6 Comparaison de modèles 209
10.6.1 Vraisemblance sous le modèle T92+CpGs pour l'alignement 1 211
10.6.2 Vraisemblance sous le modèle T92+CpGs pour l'alignement 2 212
10.6.3 Comparaison des vraisemblances obtenues sous trois modèles 213

Perspectives **215**

A Description des modèles **217**

B Identifiabilité **221**

C Bibliographie succincte sur les modèles avec dépendance **227**
C.1 Approches avec troncature de la dépendance 227
C.2 Approches sans troncature de la dépendance 229

Introduction

Cette thèse s'inscrit dans le domaine des probabilités appliquées et de la statistique mathématique. Le cœur du travail effectué concerne l'application de méthodes particulaires à des problèmes d'inférence en évolution moléculaire.

- Les méthodes particulaires [21] sont des méthodes numériques stochastiques reposant sur la représentation approchée d'une loi de probabilité par la distribution empirique associée à un grand nombre de particules.
- L'évolution moléculaire [54] est un domaine qui étudie l'évolution biologique au niveau des séquences d'ADN.

Les modèles stochastiques d'évolution de l'ADN le long d'un arbre [43] constituent l'un des outils essentiels employés en phylogénie moléculaire. Plus spécifiquement, on s'intéresse dans ce travail à des modèles stochastiques d'évolution décrivant des mutations par substitution de nucléotide, à l'exclusion des autres types de mutation existants (par exemple les mutations par insertion ou délétion). Le plus souvent, ces modèles stochastiques sont utilisés sous l'hypothèse que les différents sites évoluent de manière indépendante. Dans cette thèse au contraire, on s'intéresse à des modèles stochastiques de substitution de nucléotides avec dépendance au contexte.

Ces modèles permettent de rendre compte du fait que, pour beaucoup de séquences génomiques [20], la fréquence observée de certains dinucléotides est significativement différente de la fréquence qui serait prédite par un modèle à sites indépendants.

J'ai principalement travaillé sur la classe de modèles stochastiques avec dépendance RN95+YpR définie et étudiée dans [14]. Cette classe de modèles repose sur un renforcement des taux d'occurrence de certaines substitutions en fonction du contexte local, ce qui introduit des phénomènes de dépendance dans l'évolution des différents sites de la séquence.

La présence de dépendance – même locale – dans ce type de modèles, rend problématique l'utilisation des méthodes classiques d'inférence statistique, en raison notamment de l'impossibilité pratique de calculer la vraisemblance exacte. En effet, formellement un tel calcul se ramène à celui de l'exponentielle d'une matrice de taille $4^m \times 4^m$, où la longueur de la séquence m est par exemple de l'ordre de 10^3 ou 10^4.

L'un des buts principaux de cette thèse est d'établir des méthodes permettant d'approcher la vraisemblance exacte d'une séquence d'observations issue d'un modèle RN95+YpR et de confronter théoriquement et numériquement ces différentes méthodes.

Dans les prochains paragraphes de cette introduction, on détaille précisément les différents apports originaux de cette thèse. Ces apports s'articulent autour de cinq axes :

- la compréhension de la structure de dépendance spatiale du modèle RN95+YpR le long d'une (ou d'un ensemble de) séquence(s) observée(s),
- l'identification et l'étude de nouvelles structures de dépendance spatiales du modèle

RN95+YpR, associées à l'évolution des séquences d'ADN sur toute l'histoire évolutive,
- les aspects théoriques associés à la vraisemblance et au maximum de vraisemblance de la classe RN95+YpR,
- l'introduction de méthodes de simulation, avec en particulier des méthodes particulaires permettant d'approcher de façon consistante la vraisemblance exacte d'observations issues d'un modèle RN95+YpR,
- la mise en œuvre informatique (en C++) et l'exploitation de ces méthodes, permettant entre autres de comparer les différentes méthodes d'approximation de la vraisemblance et d'effectuer des comparaisons de modèles.

Structure des séquences observées. Les apports suivants sont reliés à la structure de dépendance spatiale le long d'une séquence observée issue d'un modèle RN95+YpR.
- *Découpage RY.* À partir des propriétés étudiées dans [14], on identifie une propriété permettant sous certaines conditions de découper l'évolution le long des séquences observées en morceaux indépendants et ainsi de simplifier le calcul de la vraisemblance. Cette propriété est appelée découpage RY.
- *Vraisemblances composites par approximations markoviennes.* Une vraisemblance composite basée sur les encodages spécifiques de la classe de modèles RN95+YpR et nommée vraisemblance composite par triplets encodés a été étudiée dans [15]. Elle permet d'estimer de façon consistante les paramètres du modèle. Dans cette thèse, on introduit de nouvelles vraisemblances composites basées sur ces encodages, les vraisemblances composites par approximations markoviennes. L'intérêt est d'avoir une quantité du même ordre de grandeur que la vraisemblance réelle. Par contre, cette troncature conduit à n'obtenir qu'une approximation (en général non consistante) de la vraisemblance réelle. La pertinence de ces quantités est liée à la dépendance présente le long d'une séquence.
- *Étude de la dépendance le long d'une séquence.* Un chapitre de cette thèse est consacré aux phénomènes de dépendance le long d'une séquence d'observations pour les modèles RN95+YpR. Des exemples, pour lesquels les calculs théoriques exacts peuvent être effectués, mettent en évidence que l'évolution le long des séquences n'est pas une chaîne de Markov (d'ordre quelconque) en général.

Structures de dépendance spatiales associées à l'évolution globale. L'étude de la dépendance le long d'une séquence d'observations montre que la structure obtenue est complexe et difficile à analyser. On cherche à disposer d'un autre point de vue, dans lequel la structure de dépendance est mieux identifiée. On considère alors l'ensemble de l'histoire évolutive, c'est-à-dire évolution le long de l'arbre depuis la séquence ancestrale jusqu'aux séquences observées. On élucide dans cette thèse trois nouvelles structures de dépendance spatiales (c'est-à-dire vis-à-vis des sites) associées aux propriétés des modèles RN95+YpR. Ces structures de dépendance markoviennes sont cachées, dans le sens où elles sont données par les histoires évolutives et pas seulement par les séquences observées.
- *Structure de champ markovien.* Pour un modèle avec dépendance aux voisins immédiats général, il est connu [27, 28] que l'évolution temporelle d'un nucléotide en un site i conditionnellement aux sites voisins ne va dépendre que des sites $(i-2, i-1, i+1, i+2)$, c'est-à-dire que l'on obtient une structure spatiale de champ markovien d'ordre deux. Dans cette thèse, on montre que la classe RN95+YpR possède une

structure de champ markovien d'ordre un. Cette structure de dépendance est rendue explicite en termes de noyaux de transitions (avec ou sans conditionnement par les séquences observées). Cette structure de champ markovien induit une structure de chaîne de Markov d'ordre un, qui n'est pas exploitable directement du fait de son caractère non explicite.

– *Structure de chaîne de Markov explicite.* En exprimant le modèle RN95+YpR différemment, on identifie une autre structure spatiale de chaîne de Markov qui permet de faire apparaître la dépendance le long des séquences observées comme une chaîne de Markov cachée. L'idée est de ne pas considérer l'évolution de chaque nucléotide, mais de regarder l'évolution de chaque dinucléotide, avec chevauchement, encodé dans un espace plus petit que $\{A, C, G, T\} \times \{A, C, G, T\}$. Cette nouvelle approche permet de faire apparaître une structure de dépendance markovienne. La structure de dépendance obtenue est explicite, c'est-à-dire que l'on sait exprimer explicitement les noyaux de transition correspondants (avec ou sans conditionnement), qui sont nécessaires à la spécification des algorithmes particulaires envisagés.
– *Structure de chaîne de Markov sur un alphabet restreint.* La dernière structure de dépendance obtenue est basée non pas sur l'alphabet $\{A, C, G, T\}$ mais sur un alphabet plus petit noté $\{R, Y\}$. À partir d'une évolution encodée dans cet alphabet restreint, on reconstruit ensuite l'évolution de la séquence non encodée conditionnellement à l'évolution effectuée dans l'alphabet $\{R, Y\}$. L'intérêt de cette structure de dépendance est que, conditionnellement à l'évolution dans l'alphabet $\{R, Y\}$, l'évolution de chaque dinucléotide (correctement encodé) devient indépendante des autres dinucléotides et peut être décrite explicitement.

Comportement asymptotique d'estimateurs. On apporte dans cette thèse les résultats théoriques suivants, liés à la vraisemblance ou au maximum de vraisemblance des séquences d'observations issues d'un modèle RN95+YpR.

– *Consistance et normalité asymptotique de l'estimateur du maximum de vraisemblance.* À partir de la structure de chaîne de Markov cachée obtenue, on établit un théorème de consistance et de normalité asymptotique du maximum de vraisemblance pour les modèles avec dépendance RN95+YpR. Ce théorème s'appuie sur le théorème de consistance et de normalité asymptotique pour les chaînes de Markov cachées établi dans [21].
– *Conditions d'identifiabilité du modèle.* À partir des vraisemblances composites par triplets encodés définies dans [15], on établit des conditions d'identifiabilité explicites du modèle RN95+YpR.
– *Estimation de la variance de l'estimateur du maximum de vraisemblance composite par triplets encodés.* Dans [15], il est mentionné que la vraisemblance composite par triplets encodés fournit un estimateur convergent des paramètres du modèle. Dans cette thèse, on construit un estimateur semi-empirique de la variance de cet estimateur, en utilisant les dérivées premières et secondes de la vraisemblance composite en chaque triplet.

Algorithmes de simulation. Les différentes structures de chaînes de Markov de l'évolution permettent de mettre en œuvre différents algorithmes de simulations.

– *Approximation de la vraisemblance à l'aide de méthodes particulaires.* La structure de chaîne de Markov explicite de l'évolution permet d'utiliser des méthodes particu-

laires pour la chaîne de Markov cachée associée. Ici un état de la chaîne de Markov cachée correspond à une évolution le long de l'arbre (c'est-à-dire de l'instant initial aux instants finaux) d'un dinucléotide encodé de façon spécifique. L'algorithme utilisé est le filtre particulaire auxiliaire, dans le cas optimal, avec ou sans rééchantillonnage. En utilisant les résultats asymptotiques existants, on obtient que ces méthodes particulaires permettent d'approcher de façon consistante [26] et asymptotiquement normale [23] la vraisemblance exacte des séquences observées issues d'un modèle RN95+YpR.

- *Simulation sans dégénérescence d'une évolution issue d'un modèle RN95+YpR conditionnée par les séquences observées.* En utilisant la structure de chaîne de Markov explicite et le découpage RY, on propose un algorithme particulaire permettant de simuler selon la loi de l'évolution conditionnée par les observations, sans problème de dégénérescence des poids (cas avec une méthode particulaire sans rééchantillonnage) et sans problème de dégénérescence de la généalogie (cas avec une méthode particulaire avec rééchantillonnage).
- *Simulation exacte de la loi stationnaire.* À l'aide de la structure de chaîne de Markov sur un alphabet restreint de l'évolution, on établit un algorithme de simulation exacte de la loi stationnaire, inspiré de l'algorithme de couplage depuis le passé de Propp-Wilson [98].

Applications. On décrit dans ce paragraphe l'implémentation et l'utilisation de méthodes de simulation particulaires approchant de façon consistante la vraisemblance exacte. La majeure partie de l'analyse a été effectuée sur des données simulées.

- *Implémentation.* Une partie de la thèse a été consacrée à élaborer et à valider des codes de simulations en C++ mettant en œuvre les algorithmes d'approximation de la vraisemblance par méthodes particulaires, avec ou sans rééchantillonnage. De plus, des codes permettant de calculer les vraisemblances composites par approximations markoviennes ont été réalisés. Ces méthodes numériques fournissent des résultats dont les points suivants décrivent plusieurs exemples d'exploitation.
- *Calculs d'approximations de la vraisemblance et comparaisons.* En utilisant l'approximation de la vraisemblance par les méthodes particulaires, il est possible de comparer la vraisemblance exacte avec la vraisemblance composite par approximation markovienne. On mesure sur des exemples typiques et atypiques l'écart obtenu entre les deux approximations.
- *Inférence de nucléotides à la racine.* Pour un modèle RN95+YpR fixé, les algorithmes implémentés permettent également d'inférer la valeur d'un site de la séquence ancestrale conditionnellement aux observations et de mettre en évidence empiriquement la portée de la dépendance de la valeur inférée vis-à-vis des nucléotides voisins.
- *Comparaison empirique d'estimateurs.* On compare aussi les estimations du maximum de vraisemblance obtenues à l'aide des méthodes particulaires avec celles obtenues par le maximum de vraisemblance composite triplets par triplets développé dans [15] (les deux méthodes étant consistantes et asymptotiquement normales).
- *Comparaison de modèles.* L'approximation de la vraisemblance exacte à l'aide des méthodes particulaires permet enfin de comparer sur des données génomiques le modèle RN95+YpR avec d'autres modèles d'évolution, par exemple le modèle réversible à sites indépendants GTR [73, 112, 121].

Organisation des chapitres.

Les différents chapitres s'organisent de la façon suivante. On décrit dans le chapitre 1 les modèles d'évolution de séquences d'ADN considérés dans la thèse. Dans le chapitre 2, on introduit les méthodes d'approximation de la vraisemblance qui vont être envisagées. Le chapitre 3 présente ensuite des encodages spécifiques aux modèles RN95+YpR. On utilise dans le chapitre 4 ces encodages pour définir différentes vraisemblances composites associées aux modèles RN95+YpR. Dans le chapitre 5, on illustre les phénomènes de dépendance possibles de la classe RN95+YpR le long d'une séquence d'observations. Le chapitre 6 a pour but d'établir différentes structures de dépendance spatiales markoviennes associées à l'évolution. On montre dans le chapitre 7 la convergence et la normalité asymptotique de l'estimateur du maximum de vraisemblance pour les modèles RN95+YpR. Dans le chapitre 8, on détaille comment les méthodes particulaires peuvent être utilisées pour la structure de chaîne de Markov explicite associée au modèle RN95+YpR. Le chapitre 9 est consacré a la description de l'implémentation d'une méthode particulaire, réalisée en C++. On regroupe dans le chapitre 10 les applications numériques développées dans cette thèse. Enfin, on établit dans l'annexe B des conditions d'identifiabilité du modèle RN95+YpR et on présente dans l'annexe C une bibliographie des différentes approches ayant été introduites pour étudier les modèles avec dépendance.

Chapitre 1

Modèles d'évolution

Ce chapitre est consacré à la description des modèles d'évolution de séquences d'ADN considérés dans la thèse, et notamment à la classe de modèles avec dépendance au contexte RN95+YpR introduite dans [14].

ADN et mutations

L'information génétique contenue dans une séquence d'ADN est décrite comme une suite finie de nucléotides [54]. À chaque nucléotide est associée une base parmi l'adénine (A), la cytosine (C), la guanine (G) et la thymine (T) et on peut donc écrire la séquence d'ADN comme une suite finie d'éléments de l'alphabet :

$$\mathcal{A} := \{A, C, G, T\}. \tag{1.1}$$

Un exemple de séquence d'ADN est donné sur la figure 1.1.

```
AAAAAAAGTCATACCTATTTAGTTTATATTCATTCTAAGGCCTCTCCTTT
```

FIGURE 1.1 – Exemple de séquence d'ADN.

Une distinction importante existe entre deux types de bases : les purines (A, G) et les pyrimidines (C, T). On note l'ensemble des éléments de $\mathcal{A}$ dont les bases correspondent à des purines (resp. des pyrimidines) par :

$$R := \{A, G\} \text{ et } Y := \{C, T\}. \tag{1.2}$$

Remarque 1.0.1. *L'espace d'états d'une séquence de longueur m est $\mathcal{A}^m$ et est de cardinal 4^m.*

Remarque 1.0.2. *On sait que l'ADN possède une structure en double hélice, constitué de deux brins complémentaires, liant une base adénine à une base thymine et une base cytosine à une base guanine. Ici, toutes les séquences considérées correspondront à un même brin d'ADN. Les liaisons chimiques présentes entre les bases consécutives de ce brin permettent ensuite de lui adjoindre une orientation : on parle du sens $5'$ vers $3'$. Ainsi, par exemple, l'écriture CG correspond à deux nucléotides à la suite le long de la séquence selon cette orientation, et non à un nucléotide sur les deux brins complémentaires.*

Notation 1.0.3. *Dans l'écriture des modèles, on adopte la notation CpG pour désigner le dinucléotide CG, ceci pour souligner l'appartenance au même brin d'ADN. On note aussi YpR l'ensemble des dinucléotides constitués d'une pyrimidine suivi d'une purine.*

On souhaite étudier l'évolution au cours du temps de l'information génétique contenue dans une séquence d'ADN. Cette information génétique peut être décrite comme une suite finie de nucléotides, qui va subir des mutations affectant un ou plusieurs nucléotides (voir le chapitre 1 de [54]). On peut citer par exemple :

- les mutations ponctuelles par substitution : un nucléotide est remplacé par un autre nucléotide,
- les mutations ponctuelles par insertion : un nucléotide est ajouté entre deux nucléotides existants,
- les mutations ponctuelles par délétion : un nucléotide existant est supprimé,
- les mutations par inversion : une portion de la séquence est renversée,
- les mutations par duplication : une portion de la séquence est répétée deux fois.

On s'intéresse dans cette thèse uniquement aux mutations par substitutions ponctuelles, en négligeant les autres. Les modèles d'évolution considérés sont donc des modèles d'évolution par substitutions ponctuelles et décrivent le comportement évolutif au cours du temps.

Dans la suite, on définit des modèles stochastiques d'évolution par substitutions ponctuelles : d'abord les modèles à sites indépendants (section 1.1) puis les modèles RN95+YpR (section 1.2). Ces évolutions sont données de séquence à séquence, c'est-à-dire depuis une séquence ancestrale jusqu'à une séquence associée au temps actuel fixé. On généralise ensuite ces modèles à l'évolution le long d'un arbre (section 1.3). Après, on discute de la pertinence des hypothèses biologiques sous-jacentes à ces modèles (section 1.4). On termine ce chapitre en fournissant un cadre général pour décrire les modèles d'évolution par substitutions ponctuelles homogène dans le temps (section 1.5).

1.1 Modèles à sites indépendants

Les modèles à sites indépendants considèrent que tous les sites évoluent indépendamment selon le même processus. De plus, l'évolution de chaque site est régie par une chaîne de Markov à temps continu homogène sur l'alphabet $\mathcal{A}$ (comme références sur les chaînes de Markov en temps continu, voir par exemple [2, 19, 31, 88]). Un modèle à sites indépendants comprend alors au plus 12 paramètres, et la matrice de taux de sauts s'écrit (avec l'ordre A, C, G, T) :

$$Q = \begin{pmatrix} \cdot & \mu_{AC} & \mu_{AG} & \mu_{AT} \\ \mu_{CA} & \cdot & \mu_{CG} & \mu_{CT} \\ \mu_{GA} & \mu_{GC} & \cdot & \mu_{GT} \\ \mu_{TA} & \mu_{TC} & \mu_{TG} & \cdot \end{pmatrix} \tag{1.3}$$

où pour $x \neq y$, les taux μ_{xy} sont strictement positifs et où les termes diagonaux sont tels que la somme de chaque ligne est nulle.

Remarque 1.1.1. *Par hypothèse sur les taux, il existe une unique loi stationnaire de ce modèle notée* $(\pi_A, \pi_C, \pi_G, \pi_T)$.

Notation 1.1.2. *On utilise les notations :* $N \to N'$ *qui se lit de* N *vers* N' *;* $N \leftrightarrow N'$ *qui se lit de* N *vers* N' *ou de* N' *vers* N.

1.1.1 Hypothèses additionnelles

Réversibilité.

Une hypothèse possible d'un modèle est la réversibilité. Elle impose que sous la loi stationnaire du modèle, on ait en moyenne autant de substitutions de N vers N' que de N' vers N pour tous $N, N' \in \mathcal{A}$. Elle est définie de la façon suivante pour le modèle à sites indépendants général décrit par la matrice Q de l'équation (1.3) :

Définition 1.1.3. *Le modèle est dit réversible si les taux de sauts et la loi stationnaire vérifient, pour tous $N, N' \in \mathcal{A}$:*

$$\pi_N \mu_{NN'} = \pi_{N'} \mu_{N'N}.$$

Symétrie des brins.

Une autre hypothèse possible est celle de symétrie des brins (voir [7]).

Définition 1.1.4. *On dit que l'hypothèse de symétrie des brins – ou* strand symmetry assumption *– est vérifiée si le même modèle peut être employé pour décrire les substitutions affectant les deux brins complémentaires de l'ADN. Dans ce cas, les taux de sauts à partir d'une séquence doivent être identiques à ceux de sa séquence complémentaire retournée*

Exemple 1.1.5. *Sous l'hypothèse de la définition 1.1.4, le taux de substitution $TGGCC \to TCGCC$ (resp. $CG \to CA$, resp. $A \to C$), doit être identique au taux de substitution $GGCCA \to GGCGA$ (resp. $CG \to TG$, resp. $T \to G$).*

En notant c la fonction qui a un nucléotide $N \in \mathcal{A}$ associe le nucléotide sur la séquence complémentaire :

$$c(A) := T\,;\; c(G) := C\,;\; c(C) := G\,;\; c(T) := A,$$

l'hypothèse de symétrie des brins correspond à avoir l'égalité pour tous N, N' entre le taux de substitution de N vers N' et le taux de substitution de $c(N)$ vers $c(N')$.

On discute de la pertinence biologique des hypothèses de réversibilité et de symétrie des brins dans la section 1.4.

Normalisation de l'évolution.

Souvent, la loi de la séquence associée à la racine est prise selon la loi stationnaire. Dans ce cas, en chaque instant, la loi d'évolution est stationnaire et le nombre moyen de substitution par unité de temps est identique. Ce nombre est en outre donné par le coefficient :

$$\beta = \sum_{N \in \mathcal{A}} \pi_N \sum_{N' \neq N} \mu_{NN'}. \tag{1.4}$$

Si on ne s'intéresse pas à la distance absolue entre la séquence ancestrale et la séquence associée au temps actuel, on peut alors imposer en divisant la matrice Q par la quantité β qu'il y ait en moyenne une substitution par unité de temps.

Dans le cas du modèle à sites indépendants général, 11 paramètres subsistent après normalisation.

1.1.2 Description des modèles T92, RN95 et GTR

De nombreux modèles inclus dans le modèle général décrit par (1.3) ont été étudiés, depuis le modèle JC69 [68] décrit par Jukes et Cantor.

Notons que biologiquement les substitutions du type $C \leftrightarrow T$ et $A \leftrightarrow G$ sont en général deux à trois fois plus fréquentes que les autres substitutions (pour les mammifères voir [109]). Un modèle convenable doit donc prendre en compte ces différences. On utilise le vocabulaire suivant, standard en évolution moléculaire :

Définition 1.1.6.

- *Une transition est une substitution $R \to R$ ou $Y \to Y$, c'est-à-dire substituant une purine à une autre purine, ou une pyrimidine à une autre pyrimidine.*
- *Une transversion est une substitution $R \to Y$ ou $Y \to R$, c'est-à-dire substituant une pyrimidine à une purine ou une purine à une pyrimidine.*

Les modèles T92, RN95 et GTR que l'on va décrire maintenant prennent en compte la différence entre taux de transitions et taux de transversions.

Modèle T92.

Le modèle T92 [110] est un modèle à sites indépendants qui tient compte des différences entre transversions et transitions à travers le paramètre $\kappa > 0$ (on multiplie le taux de substitution par κ lorsqu'il s'agit d'une transition), et des différences entre aller vers $\{C, G\}$ ou aller vers $\{A, T\}$ à travers le paramètre $\theta \in]0, 1[$. La matrice de taux de sauts associée est la suivante :

$$Q = \alpha \begin{pmatrix} . & \theta & \theta\kappa & 1-\theta \\ 1-\theta & . & \theta & (1-\theta)\kappa \\ (1-\theta)\kappa & \theta & . & 1-\theta \\ 1-\theta & \theta\kappa & \theta & . \end{pmatrix},$$

où les termes diagonaux sont tels que la somme de chaque ligne est nulle.

Remarque 1.1.7. *Le paramètre $\alpha > 0$ est ici un paramètre d'échelle qui régit la vitesse des substitutions. Comme la loi stationnaire de ce modèle est donnée par : $\left(\frac{1-\theta}{2}, \frac{\theta}{2}, \frac{\theta}{2}, \frac{1-\theta}{2}\right)$, on obtient que le coefficient de normalisation est donné par : $\beta = \left(2\theta\kappa - 2\theta^2\kappa + 1\right)\alpha$. Ainsi, en fixant $\alpha = \frac{1}{2\theta\kappa - 2\theta^2\kappa + 1}$, il n'y a pas de normalisation à faire et le modèle d'évolution dépend alors de seulement 2 paramètres.*

Modèle RN95.

Le modèle RN95 [101] n'impose qu'une condition sur le modèle à sites indépendants général : que toutes les transversions soient associées à un taux de saut qui ne dépende que du nucléotide obtenu après substitution. La matrice de taux de sauts associée s'écrit donc :

$$Q = \begin{pmatrix} . & v_C & w_G & v_T \\ v_A & . & v_G & w_T \\ w_A & v_C & . & v_T \\ v_A & w_C & v_G & . \end{pmatrix},$$

où pour $x \in \mathcal{A}$, les v_x, w_x sont strictement positifs et où les termes diagonaux sont tels que la somme de chaque ligne est nulle.

Remarque 1.1.8. *Ce modèle possède a priori* 8 *paramètres distincts où pour $x \in \mathcal{A}$, v_x correspond au taux de transversion et w_x au taux de transition vers le nucléotide x. Si on impose qu'il y ait en moyenne une substitution par unité de temps (en utilisant la normalisation (1.4)), alors le modèle n'a plus que* 7 *paramètres libres. On se réfère à [102] pour les détails (calcul de la loi stationnaire, coefficient de normalisation).*

Remarque 1.1.9. *Les modèles suivants s'incluent dans le modèle RN95 : JC69 [68], K80 [71], F81 [42], HKY85/F84 [44, 59, 72], TN93 [111], et le modèle déjà vu T92.*

Par contre, le modèle suivant GTR ne s'inclut ni inclut le modèle RN95.

Modèle GTR.

Le modèle GTR (pour *generalised time-reversible*) [73, 112, 121] est le modèle d'évolution à sites indépendants le plus général possédant la propriété de réversibilité (voir définition 1.1.3). Il est décrit par la matrice de taux de sauts :

$$Q = \alpha \begin{pmatrix} . & d\pi_C & \pi_G & b\pi_T \\ d\pi_A & . & e\pi_G & a\pi_T \\ \pi_A & e\pi_C & . & c\pi_T \\ b\pi_A & a\pi_C & c\pi_G & . \end{pmatrix},$$

où les π_x (pour $x \in \mathcal{A}$), a, b, c, d et e sont strictement positifs, et où les termes diagonaux sont tels que la somme de chaque ligne est nulle.

Dans ce modèle, $(\pi_A, \pi_C, \pi_G, \pi_T)$ est alors la loi stationnaire. De plus, ce modèle a 9 paramètres libres (8 si on normalise avec (1.4) la vitesse des substitutions).

Remarque 1.1.10. *Les modèles suivants s'incluent dans le modèle GTR : JC69, K80, F81, HKY85/F84, TN93 et T92.*

Limites des modèles à sites indépendants. On a utilisé dans cette section l'hypothèse simplificatrice que tous les sites de la séquence évoluent indépendamment selon le même processus. On cherche maintenant à introduire un phénomène de dépendance entre les sites. Le modèle RN95+YpR étudié consiste à modifier les taux de substitutions d'un site en fonction du nucléotide présent en ce site et des nucléotides présents sur les sites voisins. D'autres manières d'inclure un phénomène de dépendance sont mentionnées dans la section 1.4.2.

1.2 Le modèle RN95+YpR

L'analyse statistique de données de séquences montre [20] que pour beaucoup de génomes, les modèles à sites indépendants décrivent mal la distribution réelle de la séquence, même dans les régions non codantes. En effet, la fréquence observée de certains dinucléotides est significativement différente de la fréquence attendue dans un modèle à sites indépendants – qui est alors le produit des fréquences observées de chaque nucléotide composant le dinucléotide. Par exemple, le dinucléotide TA est sous-représenté dans presque

tous les génomes, et le dinucléotide CG est très sous-représenté chez les vertébrés [67], par rapport à un modèle où les sites sont indépendants.

Chez les mammifères, un phénomène biochimique [17] – la méthylation de la cytosine – explique des taux de substitutions des dinucléotides CG vers TG ou CA environ dix fois plus importants [85] que les taux globaux de C vers T ou de G vers A. Le cas des primates a été étudié en particulier dans [122]. Le renforcement de ces deux substitutions est appelé hypermutabilité CpG.

Toujours chez les mammifères, il existe certaines portions de séquences appelées îlots CpG où au contraire les dinucléotides CG sont en nombre plus important qu'attendu [50].

On est ainsi amené à considérer des modèles qui ne sont plus à sites indépendants. Un modèle avec dépendance aux voisins est tel que pour chaque nucléotide de la séquence, les taux de substitutions à partir de ce nucléotide dépendent non seulement de la valeur parmi A, C, G, T de ce nucléotide (cas d'un modèle à sites indépendants) mais également de la valeur des nucléotides voisins immédiats. Comme pour les modèles à sites indépendants, l'hypothèse d'homogénéité du modèle d'évolution vis-à-vis des sites de la séquence est conservée.

On considère la classe particulière de modèles RN95+YpR introduite dans [14], construite de cette manière : on se base sur un modèle à sites indépendants, sur lequel on renforce avec dépendance aux voisins immédiats uniquement certaines substitutions. En particulier, il s'agit d'une chaîne de Markov homogène à temps continu sur l'espace $\mathcal{A}^m$ des séquences d'ADN de taille m, dans laquelle les seuls taux non nuls sont ceux qui correspondent à une substitution d'une lettre.

Notation utilisée. Dans la suite, la longueur des séquences étudiées est notée m et les sites sont numérotés de 1 jusqu'à m. Pour tout site i, on dit que le site $i-1$ (resp. le site $i+1$) est le site à gauche (resp. à droite) du site i, ou encore qu'il est le site précédant (resp. le site suivant) le site i. On utilise la notation suivante pour décrire globalement la séquence d'évolution.

Notation 1.2.1. *Notation de la séquence d'évolution. La séquence d'ADN évoluant de séquence à séquence du temps initial* 0 *au temps final* T *est de longueur* m *et notée*

$$X = (X(t))_{t\in[0,T]}.$$

On distingue les différents sites en notant pour tout $t \in [0,T]$ *:*

$$X(t) := (X_1(t), \ldots, X_m(t))$$

et pour chaque site i*, on note :*

$$X_i := (X_i(t))_{t\in[0,T]}.$$

1.2.1 Définition du modèle RN95+YpR

Ce modèle à dépendance est basé sur le modèle à sites indépendants RN95 et les substitutions renforcées sont les suivantes :

- pour $x \in Y$, $x' \in Y\backslash\{x\}$ et $y \in R$, on renforce la substitution du dinucléotide xy vers $x'y$ par le taux $r_{xy\to x'y}$,
- pour $x \in Y$, $y \in R$ et $y' \in R\backslash\{y\}$, on renforce la substitution du dinucléotide xy vers xy' par le taux $r_{xy\to xy'}$.

Ainsi huit renforcements de substitutions sont ajoutés, un pour chaque transition d'un dinucléotide de YpR (on rappelle que YpR est l'ensemble $\{CA, CG, TA, TG\}$). Globalement, on a par exemple que le dinucléotide TG se substitue en TA avec le taux $w_A + r_{TG\to TA}$.

Remarque 1.2.2. *En particulier, ce modèle permet de renforcer les mutations $CG \to CA$ et $CG \to TG$. Cela entraîne une évolution qui favorise la diminution de dinucléotides CG au profit des dinucléotides CA et TG. Ce type de modèle permet donc d'inclure l'hypermutabilité CpG.*

Exemple 1.2.3. *Illustration du renforcement des taux :*

- *le taux de $ATCGT$ vers $ATCAT$ est égal à $w_A + r_{CG\to CA}$,*
- *le taux de $ATCGT$ vers $ATCTT$ est égal à v_T.*

Notation 1.2.4. *On note :*

$$\mathcal{B} = \{CA{\to}CG,\ CA{\to}TA,\ CG{\to}CA,\ CG{\to}TG,\ TA{\to}CA,\ TA{\to}TG,\ TG{\to}CG,\ TG{\to}TA\}.$$

Ainsi, un jeu de paramètres du modèle RN95+YpR est la donnée d'un 16*-uplet :*

$$(v_x, w_x, r_y; x \in \mathcal{A}, y \in \mathcal{B}).$$

On partitionne $\mathcal{B}$ en deux ensembles, suivant la position du nucléotide qui ne varie pas lors de la substitution :

$$\mathcal{B}_g = \{CA{\to}CG,\ CG{\to}CA,\ TA{\to}TG,\ TG{\to}TA\}$$

et

$$\mathcal{B}_d = \{CA{\to}TA,\ CG{\to}TG,\ TA{\to}CA,\ TG{\to}CG\}.$$

D'après la définition du modèle RN95+YpR, on peut écrire la matrice de taux de sauts instantanée de l'évolution en chaque instant. Cette matrice donne le taux de saut d'une séquence $(x_1, \dots, x_i, \dots x_m)$ vers une séquence $(x_1, \dots, x'_i, \dots x_m)$ dans laquelle seul le nucléotide au site i change.

Définition 1.2.5. *La matrice de taux de sauts instantanée d'un site $i \notin \{1, m\}$ à un instant fixé dépend à cet instant du nucléotide au site précédent (le site à gauche $i-1$) noté $g \in \mathcal{A}$ et du nucléotide au site suivant (le site à droite $i+1$) noté $d \in \mathcal{A}$. Elle est donnée*

par :

$$Q_{g,d} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} . & v_C & \begin{array}{c} w_G+ \\ r_{TA\to TG}\mathbf{1}_{g=T}+ \\ r_{CA\to CG}\mathbf{1}_{g=C} \end{array} & v_T \\ v_A & . & v_G & \begin{array}{c} w_T+ \\ r_{CA\to TA}\mathbf{1}_{d=A}+ \\ r_{CG\to TG}\mathbf{1}_{d=G} \end{array} \\ \begin{array}{c} w_A+ \\ r_{TG\to TA}\mathbf{1}_{g=T}+ \\ r_{CG\to CA}\mathbf{1}_{g=C} \end{array} & v_C & . & v_T \\ v_A & \begin{array}{c} w_C+ \\ r_{TA\to CA}\mathbf{1}_{d=A}+ \\ r_{TG\to CG}\mathbf{1}_{d=G} \end{array} & v_G & . \end{pmatrix} \end{array}.$$

Remarque 1.2.6. *La connaissance de g dans $\{\{A,G\},C,T\}$ et de d dans $\{A,G,\{C,T\}\}$ suffisent à définir la matrice $Q_{g,d}$.*

1.2.2 Gestion de la condition aux bords

L'évolution des nucléotides situés sur les bords de la séquence, c'est-à-dire en position 1 et m, n'est pas clairement définie. On étudie deux conditions aux bords :

- (Condition aux bords I) On considère que les taux de saut au site 1 (resp. le site m) ne sont pas affectés par les nucléotides qui peuvent le précéder (resp. qui peuvent le suivre).
- (Condition aux bords II) On considère que la loi de l'évolution des nucléotides aux positions $[\![1,m]\!]$ s'inclut dans la loi de l'évolution des nucléotides sur $\mathbb{Z}$ (construction dans [14]). Ainsi chaque nucléotide dépend en toute position des nucléotides le suivant et le précédant.

La condition I peut ainsi être vue comme une approximation de la condition II. On détaillera par la suite (voir section 3.4) comment se comporte l'évolution pour les deux conditions.

1.2.3 Construction de l'évolution par processus de Poisson

On décrit une construction de l'évolution issue d'un modèle RN95+YpR, alternative à celle vue dans la section 1.2.1. Cette construction est issue de la section 5.1 de [14]. Elle est basée sur la création de processus de Poisson homogènes et indépendants en chaque site, à partir desquels on définit la dynamique de l'évolution.

Cette construction à l'avantage de considérer initialement des processus indépendants, qui correspondent ou non à une substitution effective d'un nucléotide suivant le contexte aux voisins.

Pour chaque site i et pour chaque $x \in \mathcal{A}$, $y \in \mathcal{B}_g$ et $y' \in \mathcal{B}_d$, on définit indépendamment :
- $\mathcal{V}_i^x$ un processus homogène de Poisson sur $\mathbb{R}$ de taux v_x,
- $\mathcal{W}_i^x$ un processus homogène de Poisson sur $\mathbb{R}$ de taux w_x,
- $\mathcal{R}_i^y$ un processus homogène de Poisson sur $\mathbb{R}$ de taux r_y,
- $\mathcal{Q}_i^{y'}$ un processus homogène de Poisson sur $\mathbb{R}$ de taux $r_{y'}$,

On note $(X_1(t), \ldots, X_m(t))_{t\in[0,T]}$ la séquence d'évolution de l'instant initial 0 jusqu'à l'horizon de temps T, où m est la longueur de la séquence. À partir d'une séquence initiale $(x_1(0), \ldots, x_m(0)) \in \mathcal{A}^m$, on définit l'évolution de la façon suivante.

Sur l'intervalle $[0, T]$, dès qu'une sonnerie d'un des processus de Poisson se déclenche, on effectue un mouvement de la façon suivante :
- Type V. Si une horloge exponentielle associée au processus $\mathcal{V}_i^x$ se déclenche, le nucléotide au site i se substitue en x dans le cas où ce mouvement correspond à une transversion.
- Type W. Si une horloge exponentielle associée au processus $\mathcal{W}_i^x$ se déclenche, le nucléotide au site i se substitue en x dans le cas où ce mouvement correspond à une transition.
- Type R. Si une horloge exponentielle associée au processus $\mathcal{R}_i^y$ se déclenche, on note $y = NN' \to NN''$ et le nucléotide au site i se substitue en N'' dans le cas où à cette date, le nucléotide au site $i-1$ est N et le nucléotide au site i est N'.
- Type Q. Si une horloge exponentielle associée au processus $\mathcal{Q}_i^y$ se déclenche, on note $y = N'N \to N''N$ et le nucléotide au site i se substitue en N'' dans le cas où à cette date, le nucléotide au site $i+1$ est N et le nucléotide au site i est N'.

Les conditions aux bords sont gérées par une des deux conditions proposées dans la section 1.2.2. La proposition suivante est alors vérifiée (issue de [14]) :

Proposition 1.2.7. *Pour un choix fixé d'une séquence initiale et d'une condition aux bords, la dynamique définie dans cette section coïncide avec celle issue de la section 1.2.1.*

Exemple 1.2.8. *On suppose que sur l'intervalle de temps* $[0,1]$*, l'horloge associée respectivement aux processus* $\mathcal{Q}_{i-1}^{CA\to TA}$*,* $\mathcal{V}_{i-1}^{A}$*,* $\mathcal{Q}_{i}^{TG\to CG}$*,* $\mathcal{W}_{i+1}^{A}$*,* $\mathcal{R}_{i+1}^{TG\to TA}$ *s'est déclenchée une fois, les marqueurs associés étant désignés respectivement par* $Q_{i-1}^{CA\to TA}(1)$*,* $V_{i-1}^{A}(1)$*,* $Q_{i}^{TG\to CG}(1)$*,* $W_{i+1}^{A}(1)$*,* $R_{i+1}^{TG\to TA}(1)$ *(chaque marqueur contient l'instant d'émission et le nom du processus auquel il est rattaché). On suppose qu'aucune autre horloge ne s'est déclenchée. On représente schématiquement cette réalisation sur la figure 1.2.*

La dynamique dépend de la séquence initiale choisie.

Si on considère la séquence initiale $ACAG$ *et* $i = 3$*, la séquence va se substituer successivement* $ATAG$ *(à partir de* $Q_{i-1}^{CA\to TA}(1)$*),* $ATAA$ *(à partir de* $W_{i+1}^{A}(1)$*) et* $AAAA$ *(à partir de* $V_{i-1}^{A}(1)$*).*

Si on considère la séquence initiale $AGTC$ *et* $i = 3$*, la séquence ne va pas subir de substitution sur l'intervalle* $[0,1]$*.*

1.2.4 Modèles inclus dans le modèle RN95+YpR

Tous les modèles à sites indépendants Mi s'incluant dans RN95 peuvent être étendus en modèles incluant la dépendance YpR et sont notés Mi+YpR. De plus, lorsque l'on considère uniquement les renforcements $CG \to CA$ et $CG \to TG$, on note Mi+CpG le modèle associé, par exemple le modèle T92+CpG.

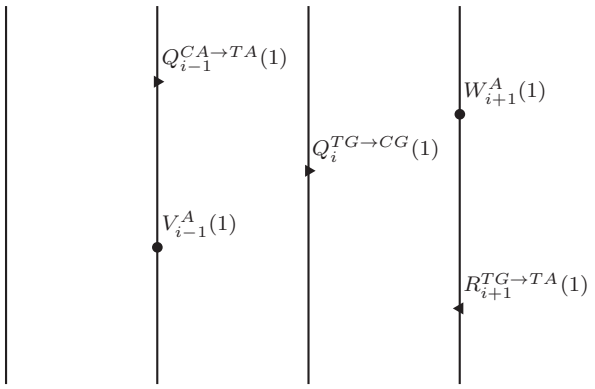


FIGURE 1.2 – Schéma de la réalisation des processus de Poisson indépendants de l'exemple 1.2.8 pour $i = 3$. Chaque segment vertical correspond à un site (de $i-2$ à $i+1$ de gauche à droite). Chaque point d'un segment correspond à un instant, de l'instant 0 (extrémité haute du segment) à l'instant 1 (extrémité basse du segment).

Dans les paragraphes suivants, on considère les modèles RN95+YpR vérifiant les propriétés additionnelles de symétrie des brins, de réversibilité et de normalisation analogues à celles énoncées dans la section 1.1.1.

Hypothèse de symétrie des brins.

Pour un modèle à sites indépendants Mi (resp. les modèles de renforcements de substitutions YpR, CpG), on appelle Mis (resp. YpRs, CpGs) le modèle associé où les conditions de l'hypothèse de symétrie des brins sont vérifiés (voir définition 1.1.4).

En particulier, CpGs correspond à avoir l'égalité entre les deux taux de renforcement $r_{CG\to CA}$ et $r_{CG\to TG}$.

Pour un modèle RN95+YpR général, l'hypothèse de symétrie des brins impose les conditions suivantes sur les paramètres du modèle RN95s+YpRs :

$$v_A = v_T \,;\, v_C = v_G \,;\, w_A = w_T \,;\, w_C = w_G,$$

$$r_{CA\to CG} = r_{TG\to CG} \,;\, r_{CA\to TA} = r_{TG\to TA} \,;\, r_{CG\to CA} = r_{CG\to TG} \,;\, r_{TA\to TG} = r_{TA\to CA}.$$

Réversibilité.

On sait que pour tout modèle RN95+YpR évoluant sur m sites et l'une des conditions aux bords, le processus markovien associé admet une unique loi stationnaire $(\pi_\varsigma)_{\varsigma\in\mathcal{A}^m}$. La définition 1.1.3 de réversibilité d'un modèle à site indépendant se généralise pour un modèle à dépendance de la façon suivante.

Définition 1.2.9. *Le modèle est dit réversible si les taux de sauts et la loi stationnaire vérifient, pour tous* $\varsigma_1, \varsigma_2 \in \mathcal{A}^m$ *:*

$$\pi_{\varsigma_1}\mu_{\varsigma_1\varsigma_2} = \pi_{\varsigma_2}\mu_{\varsigma_2\varsigma_1}.$$

La propriété suivante indique que l'hypothèse de réversibilité est très restrictive dans le cadre des modèles RN95+YpR.

Propriété 1.2.10. *Les modèles* Mi*+YpR sont en général non réversibles, même si* Mi *est lui-même réversible (voir appendice de [15]). Par exemple, même JC69+CpGs n'est pas un modèle réversible.*

Normalisation.

On suppose que la loi de la séquence associée à la racine est prise selon la loi stationnaire. Comme pour les modèles à sites indépendants, la loi d'évolution est alors stationnaire et le nombre moyen de substitution par unité de temps est donné par le coefficient :

$$\beta = \sum_{\varsigma_1 \in \mathcal{A}^m} \pi_{\varsigma_1} \sum_{\varsigma_2 \neq \varsigma_1} \mu_{\varsigma_1 \varsigma_2}. \tag{1.5}$$

On peut alors imposer en divisant la matrice Q par la quantité β qu'il y ait en moyenne une substitution par unité de temps.

1.3 Évolution RN95+YpR sur un arbre

On a défini dans la section 1.2 des modèles d'évolution par substitutions ponctuelles permettant de décrire l'évolution de séquence à séquence, c'est-à-dire depuis une séquence ancestrale jusqu'à une séquence associée au temps actuel. On considère maintenant plusieurs séquences associées au temps actuel et on cherche à représenter la proximité entre ces différentes séquences d'ADN observées. On va alors représenter les relations de proximité par un arbre phylogénétique enraciné et non enraciné (section 1.3.1), avant de définir les modèles d'évolution sur ces arbres à partir des modèles d'évolution décrits de séquence à séquence (section 1.3.2).

1.3.1 Arbre phylogénétique

Les relations de proximité entre les séquences observées vont être représentées par une structure d'arbre phylogénétique.

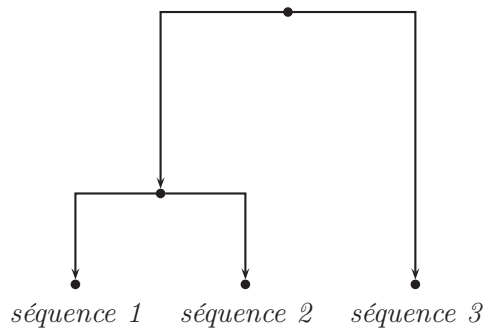


FIGURE 1.3 – Exemple d'arbre phylogénétique enraciné à trois séquences.

Dans ce qui suit, nous donnons une formulation mathématique précise de deux structures d'arbre phylogénétique. Pour des compléments sur ces définitions et sur le comptage des arbres phylogénétiques enracinés ou non, on peut se reporter aux chapitres 1 et 4 de [18], à à [41] et au chapitre 5 de [54]. La première définition correspond à celle d'un arbre phylogénétique binaire enraciné. Un exemple d'un tel arbre est représenté sur la figure 1.3.

Définition 1.3.1. *Un arbre phylogénétique binaire enraciné est un graphe orienté acyclique composé d'une unique racine et d'au moins deux sommets, tel que tous les nœuds sauf la racine possèdent un unique parent et tel que chaque nœud ait* 0 *ou* 2 *fils. L'orientation est définie à partir de la racine. Les nœuds avec* 0 *fils sont appelés feuilles de l'arbre.*

Parfois, on ne sait pas placer la racine et on définit donc une structure de parenté où la racine n'est pas fixée. Cela conduit à la structure d'arbre phylogénétique binaire non enraciné, qui est la seconde structure d'arbre phylogénétique décrite.

Définition 1.3.2. *Un arbre phylogénétique non enraciné est un graphe non orienté, acyclique, connexe, tel que chaque nœud ait un degré d'incidence de* 1 *ou* 3*. Les nœuds qui ont un degré* 1 *sont appelés feuilles de l'arbre.*

Remarque 1.3.3. *Lien avec les relations de parenté entre les espèces. Si chaque séquence est associée à une espèce, les feuilles de l'arbre correspondent alors aux espèces observées à l'heure actuelle et chaque nœud à l'ancêtre commun de ses descendants.*

Si l'arbre phylogénétique est enraciné, on identifie l'ancêtre commun de toutes les espèces considérées à la racine de l'arbre.

Convention 1.3.4. *Pour alléger la terminologie, on utilise dans la suite le terme arbre (resp. arbre non enraciné) pour un arbre phylogénétique binaire enraciné (resp. un arbre phylogénétique binaire non enraciné).*

Remarque 1.3.5.

- *On associe à chaque arbre un unique arbre non enraciné en considérant v_1, v_2 les fils de la racine, en enlevant l'orientation, la racine et ses arêtes incidentes, et en rajoutant l'arête (non orientée) $\{v_1, v_2\}$.*
- *À partir d'un arbre non enraciné comportant k arêtes, il existe k façons de lui associer un arbre. En effet, ayant considéré une arête $\{v_1, v_2\}$ parmi les k possibles, on retire cette arête, on ajoute un nœud r et les arêtes $\{r, v_1\}$ et $\{r, v_2\}$. On définit ensuite l'orientation à partir de la racine r.*

On remarque alors que la notion de feuille reste inchangée par ces opérations.

On regardera dans l'annexe B l'identifiabilité d'un arbre, qui se fera à équivalence près selon la définition suivante (voir aussi [24] section 3).

Définition 1.3.6. *Soit $T_1 = (V_1, E_1)$ et $T_2 = (V_2, E_2)$ deux arbres (ou deux arbres non enracinés) qui possèdent les même feuilles. On dit que T_1 et T_2 sont équivalents s'il existe une fonction bijective $\gamma : V_1 \to V_2$ vérifiant $\gamma(v) = v$ pour toute feuille et $E_2 = \{\{\gamma(r), \gamma(s)\}; \{r, s\} \in E_1\}$. Cela signifie que deux arbres sont équivalents s'ils sont égaux à réétiquetage près des nœuds qui ne sont pas des feuilles.*

On dit alors que les arbres T_1 et T_2 ont la même topologie.

On représente sur la figure 1.4 les topologies possibles pour un arbre à deux feuilles et pour un arbre à trois feuilles. On remarque que l'équivalence de topologies est une notion plus fine que celle d'isomorphisme de graphes : par exemple pour les arbres à quatre feuilles $\{f1, f2, f3, f4\}$, il n'existe que deux graphes non isomorphes mais 15 topologies différentes (voir figure 1.5). Sur la figure 1.6 on représente les arbres non enracinés non isomorphes à quatre feuilles ou moins.

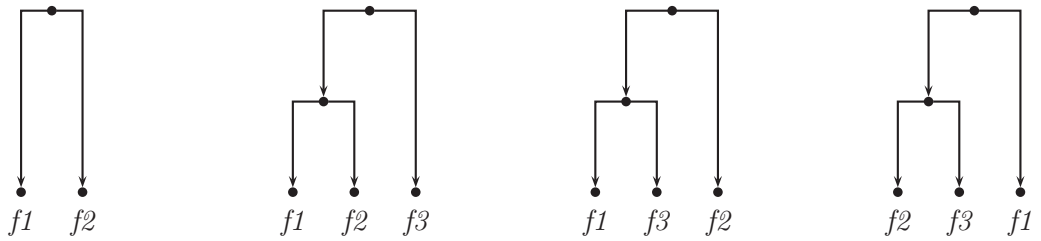


FIGURE 1.4 – Topologies possibles pour les arbres à deux feuilles $\{f1, f2\}$ ou trois feuilles $\{f1, f2, f3\}$.

FIGURE 1.5 – Représentation des deux classes d'arbres non isomorphes à quatre feuilles. La première structure comprend douze topologies différentes et la deuxième trois, suivant l'ordre choisi pour les feuilles.

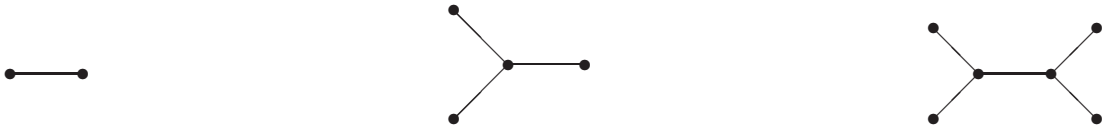


FIGURE 1.6 – Représentation des classes d'arbres non enracinés non isomorphes à quatre feuilles ou moins. Les deux premières structures comprennent une seule topologie et la troisième trois, suivant l'ordre choisi pour les feuilles.

1.3.2 Écriture globale du modèle

On va décrire le modèle d'évolution dans le cas où l'arbre phylogénétique est enraciné, avant d'en déduire le modèle d'évolution sur l'arbre non enraciné associé.

Description de l'évolution sur un arbre enraciné. On considère M un modèle d'évolution par substitutions ponctuelles inclus dans la classe RN95+YpR et évoluant sur l'espace des séquences S (typiquement $S = \mathcal{A}^m$).

On choisit T constitué d'une topologie d'arbre phylogénétique enraciné et des différentes longueurs de branches. Plus précisément, on associe à chaque arête orientée (v_1, v_2) de l'arbre T une longueur de branche $t_{(v_1,v_2)}$.

On fixe enfin une loi à la racine R définie sur S. On considère le cas où R est une séquence fixée, le cas où R est une approximation de la loi stationnaire du modèle M et le cas où R est la loi stationnaire de M.

Définition 1.3.7. *On écrit le modèle complet comme :*

$$\lambda = (\mathsf{R}, \mathsf{T}, \mathsf{M}).$$

Il est décrit de la façon suivante :

- *la séquence associée au nœud à la racine suit la loi* R*,*
- *pour chaque arête orientée* (v_1, v_2) *de longueur* $t_{(v_1,v_2)}$ *de* T*, l'évolution le long de cette arête sachant la séquence* $x(v_1)$ *associée au nœud* v_1 *suit le modèle* M *d'état initial* $x(v_1)$*,*
- *à chaque nœud intermédiaire de* T*, on définit l'état initial des deux arêtes issues de ce nœud comme l'état final de l'arête arrivant sur ce nœud,*
- *conditionnellement à leurs états initiaux, les évolutions des arêtes sont indépendantes.*

On donne dans l'annexe A la liste des modèles qui seront utilisés.

Notation 1.3.8. *Les séquences associées aux feuilles de l'arbre sont aussi appelées* observations.

Par abus de notation et pour simplifier l'écriture des observations sur un arbre, on note $X(T)$ *l'ensemble des séquences observées (aux instants finaux) et* $X_i(T)$ *l'ensemble des nucléotides observés au site* i *(aux instants finaux).*

Remarque 1.3.9. *La définition 1.3.7 vérifie l'hypothèse d'homogénéité vis-à-vis de la position dans l'arbre, c'est-à-dire que la séquence évolue selon le même modèle sur chaque arête de l'arbre ; ainsi le nombre de substitutions le long d'une arête est en moyenne proportionnel à la longueur de l'arête.*

Dans le cas où on ne s'intéresse qu'aux longueurs relatives des branches entre elles, il suffit de changer les longueurs de branches de l'arbre pour considérer que certaines séquences ont évolué plus rapidement que d'autres depuis la racine. En pratique, les distances temporelles entre la racine et les différentes feuilles seront ici différentes.

Remarque 1.3.10. *La remarque associée à l'équation (1.5) peut se reformuler sur un arbre de la façon suivante. Si la loi de la séquence associée à la racine est prise selon la loi stationnaire et si on ne s'intéresse qu'aux longueurs relatives des branches entre elles, on peut imposer en divisant la matrice de taux de sauts* Q *par le coefficient* β *(défini dans l'équation (1.5)) qu'il y ait en moyenne une substitution par unité de temps sur chaque branche de l'arbre.*

Modèle phylogénétique non enraciné associé à un modèle phylogénétique. On cherche dans ce paragraphe à déduire le modèle d'évolution sur l'arbre non enraciné associé

à un modèle global λ sur un arbre enraciné. En particulier, on cherche à exprimer les probabilités de transition sur chaque arête en fonction des probabilités de transition du modèle enraciné. L'intérêt de décrire ce lien réside dans la possibilité d'utiliser ensuite le théorème d'identifiabilité de Chang sur les arbres non enracinés (voir [24], et le théorème B.0.5 de l'annexe B consacrée à montrer l'identifiabilité du modèle).

Définition 1.3.11. *Pour $\lambda = (\mathsf{R}, \mathsf{T}, \mathsf{M})$ un modèle d'évolution sur un arbre enraciné, on note T_n l'arbre non enraciné associé à T (voir la remarque 1.3.5) et Q la matrice de taux de sauts d'espace d'états l'ensemble des séquences de longueur m fixée.*

On définit le modèle global non enraciné $\tilde{\lambda} = (\mathsf{R}, \mathsf{T}_n, \mathsf{M})$ par les transitions et les lois des nœuds issues du modèle enraciné.

On regarde maintenant explicitement la matrice de transition dans le modèle $\tilde{\lambda}$ des arêtes de l'arbre, et en particulier l'arête qui contenait initialement la racine.

Propriété 1.3.12. *Soit $\{v_1, v_2\}$ les deux sommets issus de la racine dans l'arbre T. On note t_1, t_2 la distance à la racine de respectivement v_1, v_2. On note également p le vecteur de probabilités associé à la loi à la racine R.*

La matrice de transition P_{v_1,v_2} de v_1 vers v_2 s'exprime pour toutes séquences $x, z \in S$ par :

$$P_{v_1,v_2}(x,z) = \sum_y p(y) e^{t_1 Q}(y,x) e^{t_2 Q}(y,z) / \sum_y p(y) e^{t_1 Q}(y,x). \tag{1.6}$$

Pour (v_3, v_4) une arête orientée de l'arbre T de longueur t_3 et qui ne contient pas la racine, la matrice de transition de v_3 vers v_4 sur l'arbre non enraciné s'exprime pour tous x, z par :

$$P_{v_3,v_4}(x,z) = e^{t_3 Q}(x,z).$$

Démonstration. La deuxième propriété vient de la définition de la matrice de transition associée à Q (voir aussi la propriété 1.5.5). Pour la première propriété, on note $X(v_1)$ (resp. $X(v_2)$, $X(r)$) la séquence associée au nœud v_1 (resp. v_2, r). On écrit pour $x, z \in S$:

$$P := P(X(v_2) = z | X(v_1) = x) = \frac{P(X(v_2) = z, X(v_1) = x)}{P(X(v_1) = x)}.$$

On en déduit en utilisant le caractère markovien (dans le temps) de l'évolution :

$$P = \frac{\sum_y P(X(v_2) = z | X(r) = y) P(X(v_1) = x | X(r) = y) P(X(r) = y)}{\sum_y P(X(v_1) = x | X(r) = y) P(X(r) = y)},$$

et on conclut en utilisant la définition de la matrice de transition associée à Q. □

Remarque 1.3.13. *Supposer la réversibilité du modèle M permettrait de simplifier la matrice de transition exprimée en (1.6) dans le cas où la loi à la racine est stationnaire (voir par exemple le principe de poulie dans [42]), mais cette hypothèse est trop restrictive pour la classe de modèles RN95+YpR puisque même les modèles JC69+CpGs ne sont pas réversibles (voir propriété 1.2.10).*

1.4 Pertinence biologique des hypothèses

Pour chaque hypothèse du modèle RN95+YpR, on cherche à commenter sa pertinence biologique et son cadre d'utilisation, avant donner des références sur des modèles alternatifs se passant de cette hypothèse. Des références complémentaires sont disponibles dans [45, 54].

Notons que des approches alternatives pour étudier les phénomènes de dépendance au contexte sont regroupées dans l'annexe C.

1.4.1 Alignement de séquences

Lorsque l'on considère des séquences d'ADN homologues de différentes espèces, ces séquences ont subi au cours du temps d'autres mutations que celles par substitutions ponctuelles – par exemple des insertions et délétions. Le but de l'alignement de séquences consiste à identifier les sites correspondant aux insertions et délétions de nucléotides pour ensuite positionner les séquences les unes en-dessous des autres de façon à faire ressortir et aligner les régions homologues. Cet alignement introduit en général des trous dans la séquence, notés par un tiret - et appelées lacunes, aux sites ou une insertion ou une délétion s'est produite. Un exemple d'alignement de deux séquences d'ADN est donné sur la figure 1.7.

```
GATACCGGACAG
GATAACCGGATG
```
(a) Non alignées.

```
GATA-CCGGACAG
GATAACCGGAT-G
```
(b) Alignement possible.

FIGURE 1.7 – Illustration d'un alignement possible de deux séquences d'ADN homologues issues de deux espèces.

De nombreuses méthodes d'alignements existent et ont été étudiées (voir [54] chapitre 3, *Alignement of Nucleotide and Amino Acid Sequences*). Les séquences utilisées par la suite seront supposées alignées et sans lacunes – notamment pour les séquences génomiques de la section 10.6.

Modèles avec indel. Certains modèles intègrent directement la possibilité d'obtenir des insertions ou des délétions, ponctuelles ou non. On parle alors de modèles avec indel (pour une description et une mise en œuvre de certains de ces modèles, voir par exemple [81] et les références de cet article).

1.4.2 Hypothèses d'homogénéité

Homogénéité vis-à-vis des sites. Le modèle RN95+YpR est homogène vis-à-vis des sites dans le sens où les taux de substitutions utilisés pour décrire ce modèle ne dépendent pas directement de la position i du site considéré.

Homogénéité globale. L'hypothèse d'homogénéité vis-à-vis des sites n'est pas vérifiée en général si on considère une séquence de taille importante, dans le sens où même si on suppose que les taux de mutations sont les mêmes à l'échelle des individus, la probabilité de fixation de cette mutation dépend du caractère positif, neutre ou négatif de cette mutation (voir [54] chapitre 4, *Causes of Variation in Substitution Rates*). On considère

alors des portions réduites dont on estime que l'homogénéité globale est vérifiée, appelées isochores (voir [54] chapitre 8, *Compositional Organization of the Vertebrate Genome*).

Pour une discussion sur l'hypothèse de neutralité et en particulier sur les liens avec la sélection naturelle et le mutationnisme, on se reporte à l'article [86].

Homogénéité locale. De façon locale, l'homogénéité vis-à-vis des sites n'est pas toujours vérifiée, par exemple lorsque la séquence est issue d'une région codante. Dans ce cas, chaque triplet de nucléotides va être traduit en un acide aminé, qui va définir la structure de la protéine. Si une substitution se produit sur ce triplet, alors l'acide aminé peut être le même (mutation silencieuse) ou différent. Par exemple, les triplets GTT, GTC, GTA et GTG se traduisent en l'acide aminé appelé valine, mais ATT se traduit en isoleucine.

Pour prendre en compte cette non homogénéité locale vis-à-vis des sites, des modèles ont introduit des vitesses d'évolution différentes suivant le site considéré (voir [121] et les références associées).

Homogénéité dans le temps et vis-à-vis de la position dans l'arbre. Le modèle RN95+YpR fait l'hypothèse que l'évolution est homogène dans le temps et vis-à-vis de la position dans l'arbre, c'est-à-dire que sur chaque branche et qu'en chaque instant le modèle considéré est le même.

Par contre, on n'a pas supposé de condition particulière sur les longueurs de chaque branche. Dans le cas où on ne s'intéresse qu'aux longueurs relatives des branches entre elles, changer ces longueurs correspond alors à modifier la vitesse d'évolution sur chaque branche.

Dans le cas où on s'intéresse aux valeurs absolues de chaque branche de l'arbre, on ne peut pas modifier la vitesse d'évolution de chaque branche et on est dans le cadre de l'hypothèse de l'*horloge moléculaire* (voir chapitre 4 de [54] paragraphe *Molecular Clocks*). Cette hypothèse, énoncée dans [123] en 1965, exprime l'existence d'une horloge moléculaire universelle, c'est-à-dire de taux de mutations constants au cours du temps et pour toutes les espèces. Cette hypothèse a été ensuite infirmée [29, 52] mais localement (c'est-à-dire pour un ensemble d'espèces proches), des horloges locales peuvent exister [89].

Des modèles incluent directement la possibilité de changement des taux de substitutions dans le temps ou suivant la branche considérée, par exemple [49, 113] (dans [49] une non homogénéité locale vis-à-vis des sites est également présente, de la même manière que dans [121]).

1.4.3 Hypothèses supplémentaires

Dans les définitions 1.2.9 et 1.1.4, deux hypothèses additionnelles ont été introduites : l'hypothèse de réversibilité et l'hypothèse de symétrie des brins. Nous allons maintenant donner des références sur les observations biologiques qui nous conduisent à considérer ou non ces hypothèses.

Hypothèse de réversibilité. L'hypothèse de réversibilité d'un modèle signifie que sous la loi stationnaire du modèle, l'évolution a la même loi allant du temps 0 au temps T ou allant du temps T au temps 0.

Une manière d'étudier cette hypothèse est donnée dans [120], où le modèle réversible à sites indépendants le plus général GTR est comparé avec le modèle à sites indépendants le plus général à 12 paramètres. Sur deux jeux de données étudiées, il est observé que les vraisemblances obtenues ne sont pas significativement meilleures avec le modèle à 12 paramètres.

Toutefois, dans [106, 107], des indicateurs permettant de quantifier le degré de non réversibilité sont définis, pour des modèles d'évolution sans dépendance et des modèles prenant en compte l'hypermutabilité CpG. Ces indicateurs sont ensuite utilisés sur des séquences génomiques de deux espèces (Drosophila simulans et l'homme). Pour ces deux espèces, les indicateurs permettent de rejeter l'hypothèse de réversibilité de l'évolution.

Hypothèse de symétrie des brins. L'hypothèse de symétrie des brins signifie que le même modèle d'évolution décrit les deux brins complémentaires de l'ADN. Une explication et une vérification de cette hypothèse a été effectuée dans [12] jusqu'à des séquences de longueur 9. Cette hypothèse n'est toutefois pas universelle, des contres-exemples étant données pour certaines espèces [46, 78].

1.5 Modèles d'évolution généraux

On introduit dans cette section un formalisme général pour les modèles d'évolution par substitutions ponctuelles de séquence à séquence décrits par un processus de Markov en temps continu homogène. Cela signifie que l'évolution en temps est markovienne sur l'espace des séquences et que le taux de saut d'une séquence à une autre séquence ne dépend pas de l'instant considéré (cette hypothèse est discutée dans la section 1.4.2). De plus, les seules substitutions autorisées sont des changements d'une lettre de la séquence.

En particulier, les modèles présentés dans les sections 1.1 et 1.2 s'incluent dans ce formalisme.

1.5.1 Description de la dynamique

On considère l'espace des séquences S fini, qui correspond typiquement à l'ensemble $\mathcal{A}^m$ des séquences d'ADN de taille m.

Comme on impose l'hypothèse d'une évolution markovienne homogène dans le temps, la dynamique est définie par les taux de sauts d'une séquence à une autre. Pour $\varsigma_1 \neq \varsigma_2 \in S$, on appelle $\mu_{\varsigma_1\varsigma_2}$ le taux de saut de la séquence ς_1 vers la séquence ς_2. Pour chaque $\varsigma_1 \in S$, on note :

$$\mu_{\varsigma_1} = \sum_{\varsigma_2;\varsigma_2\neq\varsigma_1} \mu_{\varsigma_1\varsigma_2} \quad \text{et} \quad \mu_{\varsigma_1\varsigma_1} = -\mu_{\varsigma_1}.$$

On définit aussi la matrice de taux de sauts : $Q = (\mu_{\varsigma_1\varsigma_2})_{\varsigma_1,\varsigma_2\in S}$ d'une séquence ς_1 vers une séquence ς_2. Pour alléger les notations, pour chaque séquence ς_1 on note également $Q(\varsigma_1) := -Q(\varsigma_1, \varsigma_1) \geq 0$.

Hypothèse 1.5.1. *On rappelle que comme on ne considère que des substitutions ponctuelles, si ς_1 et ς_2 diffèrent de plus d'un nucléotide alors : $Q(\varsigma_1, \varsigma_2) = 0$.*

On suppose aussi, sauf mention explicite du contraire, que lorsque ς_1 et ς_2 diffèrent d'exactement un nucléotide alors : $Q(\varsigma_1, \varsigma_2) > 0$. Cela assure en particulier l'irréductibilité et l'apériodicité de la matrice Q.

1.5.2 Espace d'états et description d'une évolution

On souhaite décrire l'espace des évolutions E d'une séquence de longueur m d'un instant initial 0 à un instant final T. À un instant fixé, chaque séquence est un élément de $\mathcal{A}^m$. De plus, les évolutions sont càdlàg (continues à droite avec limite à gauche) avec un nombre fini de sauts. Ainsi, on utilise les notations suivantes pour caractériser les évolutions de E.

Notation 1.5.2. *Chaque évolution $x \in \mathsf{E}$ s'écrit :*

$$x = (l, (0 = t_0 < t_1 < \ldots < t_l < T), (x(t_0), \ldots, x(t_l))),$$

où :
- *$l \in \mathbb{N}$ est le nombre de sauts lors de l'évolution,*
- *$0 = t_0 < t_1 < \ldots < t_l < T$ sont les instants de sauts et*
- *$(x(t_0), \ldots, x(t_l)) \in (\mathcal{A}^m)^{l+1}$ sont les valeurs des séquences en t_0 et après chaque saut.*

Remarque 1.5.3. *Les séquences $(x(t_0), \ldots, x(t_l))$ ne dépendent pas directement des instants $t_0, \ldots, t_l$, mais on utilise cette notation puisque cet ensemble $(x(t_0), \ldots, x(t_l))$ correspond à la donnée des séquences aux instants de sauts $t_0, \ldots, t_l$.*

On construit maintenant E explicitement et on le munit d'une tribu et d'une mesure.

Mesure sur les simplexes. On pose pour $l \in \mathbb{N}$:

$$\mathsf{F}_l = \{(t_0, \ldots, t_l); 0 = t_0 < t_1 < \ldots < t_l < T\}.$$

Pour $l \geq 1$, chaque élément de F_l est vu comme un élément $(t_1, \ldots, t_l)$ de $\mathbb{R}^l$ et on définit μ_l la mesure de Lebesgue sur la tribu borélienne $\mathcal{F}_l$ associée à F_l. Pour $l = 0$, on définit μ_0 par $\mu_0(\mathsf{F}_0) = 1$.

Mesure sur les séquences. Pour $l \in \mathbb{N}$, on pose δ_l la mesure discrète sur $(\mathcal{A}^m)^{l+1}$ définie pour chaque $l+1$-uplets $(x(t_0), \ldots, x(t_l))$ de séquences de longueurs m par :

$$\delta_l(\{(x(t_0), \ldots, x(t_l))\}) = 1$$

si pour chaque $k \in [\![0, l-1]\!]$, la séquence $x(t_{k+1})$ ne diffère de $x(t_k)$ qu'en un seul site (en particulier $x(t_{k+1}) \neq x(t_k)$), et par 0 sinon.

Mesure sur les évolutions. On pose enfin (avec $\amalg$ désignant l'union disjointe) :

$$\mathsf{E} = \amalg_{l \in \mathbb{N}} \left(\mathsf{F}_l \times (\mathcal{A}^m)^{l+1} \right),$$

et μ la mesure produit associée. La mesure μ est finie puisque pour chaque $l \in \mathbb{N}$, on a $\delta_l((\mathcal{A}^m)^{l+1}) = 4^m (3m)^l$ (4^m correspond au choix de la séquence initiale, puis à chaque instant de saut on dispose de $3m$ possibilités pour la nouvelle séquence, suivant le site substitué et sa nouvelle valeur), $\mu_l(\mathsf{F}_l) = \frac{T^l}{l!}$ et donc $\mu(\mathsf{E}) = 4^m e^{3mT}$.

1.5.3 Densité vis-à-vis de la mesure μ

On montre que chaque évolution admet une densité vis-à-vis de la mesure μ.

Proposition 1.5.4. *Pour une évolution*

$$x = (l, (0 = t_0 < t_1 < \ldots < t_l < T), (x(t_0), \ldots, x(t_l)))$$

régie par la matrice Q de taux de sauts sur les séquences et par la condition initiale $x(0) \in \mathcal{A}^m$, on exprime la densité $f := f_{x(0)}$ de l'évolution x par rapport à la mesure μ par :

$$f(x) := f_{x(0)}(x) := \left[\prod_{k=0}^{l-1} e^{-(t_{k+1}-t_k)Q(x(t_k))} Q(x(t_k), x(t_{k+1}))\right] e^{-(T-t_l)Q(x(t_l))}.$$

Démonstration. Dans le cas où il existe $k \in [\![0, l-1]\!]$ tel que $Q(x(t_k), x(t_{k+1})) = 0$, alors la proposition est vérifiée. Sinon, on utilise les résultats du chapitre 6 de [31], en particulier les théorèmes 1.2 et 1.3.

Pour chaque $k \in [\![0, l-1]\!]$, la probabilité d'obtenir le premier changement après l'instant t_k suit une loi exponentielle de paramètre Q. La probabilité que le premier changement soit dans l'intervalle $[t_{k+1} - \varepsilon_{k+1}/2, t_{k+1} + \varepsilon_{k+1}/2]$ de longueur $\varepsilon_{k+1} > 0$ est donc donnée par :

$$e^{-((t_{k+1}-\varepsilon_{k+1}/2)-t_k)Q(x(t_k))}(1 - e^{-Q(x(t_k))\varepsilon_{k+1}}) \sim e^{-(t_{k+1}-t_k)Q(x(t_k))} Q(x(t_k))\varepsilon_{k+1}.$$

De plus, la probabilité d'effectuer à cette instant de changement une transition de $x(t_k)$ vers $x(t_{k+1})$ est donnée par $\frac{Q(x(t_k),x(t_{k+1}))}{Q(x(t_k))}$.

Enfin, en utilisant que la loi exponentielle est sans mémoire, on obtient que la probabilité à partir de l'état initial $x(t_0)$ d'obtenir $l-1$ changements dans les intervalles $[t_{k+1} - \varepsilon_{k+1}/2, t_{k+1} + \varepsilon_{k+1}/2]$ (pour $k \in [\![0, l-1]\!]$), et de transitions $(x(t_1), \ldots, x(t_l))$ est donnée par une probabilité équivalente à :

$$f_{x(0)}(x)\varepsilon_1 \ldots \varepsilon_l.$$

On en déduit l'existence d'une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^l$, donnée par $f_{x(0)}$. □

1.5.4 Propriétés standards des chaînes de Markov en temps continu

On récapitule maintenant des propriétés standards des chaînes de Markov en temps continu sur un espace d'états fini. Les probabilités d'être dans chacun des états $\varsigma \in S$ au temps t sont données par le vecteur ligne :

$$P(t) = (p_\varsigma(t))_{\varsigma \in S}.$$

La propriété suivante indique que l'on déduit l'évolution du processus de la matrice de taux de sauts.

Propriété 1.5.5. *Pour tout temps t, l'évolution vérifie : $P'(t) = P(t)Q$. En se donnant $P(0)$ condition initiale, l'évolution est ensuite décrite entièrement par :*

$$P(t) = P(0)e^{tQ}.$$

On appelle e^{tQ} la matrice de transition associée à la matrice Q. Un élément de cette matrice est appelé taux de substitution.

On utilise aussi la définition suivante dans le chapitre 6, permettant de décrire les taux de sauts instantanés :

Définition 1.5.6. *Les matrices de transitions instantanées associées à une matrice de taux de sauts Q sont données pour tout $\varepsilon > 0$ par :*

$$q^{\varepsilon} = I + \varepsilon Q.$$

On a alors pour tout $t > 0$:

$$P(t+\varepsilon) = P(t)(q^{\varepsilon} + o(\varepsilon)).$$

Comme la matrice Q est choisie irréductible (d'après l'hypothèse 1.5.1) et que l'espace d'états est fini, cela assure qu'il existe une unique probabilité stationnaire $\Pi = (\pi_{\varsigma 1})_{\varsigma 1 \in S}$.

Remarque 1.5.7. *Les propriétés additionnelles de réversibilité, de symétrie des brins et de normalisation énoncées dans la section 1.2.4 s'étendent dans ce contexte général de la même manière.*

Dans le modèle le plus général, l'évolution est entièrement déterminée par la donnée des taux de sauts instantanés d'une séquence à une autre séquence. Néanmoins, le nombre de paramètres à estimer du modèle augmente exponentiellement en la taille de la séquence : pour une taille de séquence m, il est de $4^m 3m$ (puisque pour chaque séquence, le nucléotide présent en chaque site peut se substituer en un autre nucléotide parmi 3 possibilités). C'est pourquoi ce modèle général n'est pas utilisé et on le simplifie en considérant un nombre de paramètres à estimer qui ne dépend pas de la taille de la séquence, à travers les modèles à sites indépendants et les modèles RN95+YpR.

Chapitre 2

Vraisemblances pour les modèles RN95+YpR

Dans le chapitre 1, on a décrit différents modèles d'évolution de séquences d'ADN par substitutions ponctuelles : des modèles à sites indépendants – comme les modèles T92 et GTR – et des modèles appartenant à la classe RN95+YpR.

L'utilisation de ces modèles à des données génomiques suppose que l'on soit capable de faire de l'inférence statistique. Cela consiste notamment à être capable d'estimer les différents paramètres du modèle, comme les paramètres de taux de sauts ou la topologie et les différentes longueurs de branches de l'arbre. Cela consiste aussi à pouvoir comparer les différents modèles d'évolution étudiés.

Dans ce cadre, la méthode la plus classique d'inférence est celle du maximum de vraisemblance. À partir d'une classe de modèle $\lambda = (\lambda_\theta)_{\theta\in\Theta}$ paramétrée par un ensemble Θ et des séquences observées $x_{1:m}(T)$ (voir la notation 1.3.8) issues d'un modèle λ_{θ_0}, elle repose sur l'estimation de θ_0 par un paramètre $\hat{\theta}_m \in \Theta$ maximisant la fonction de vraisemblance définie par :

$$L_m(\theta; x_{1:m}(T)) := P(x_{1:m}(T)|\lambda_\theta).$$

On souhaite calculer effectivement la fonction de vraisemblance. De plus, on cherche à établir l'existence d'un maximum de vraisemblance, à le calculer, ainsi que des propriétés asymptotiques de convergence et de normalité asymptotique de celui-ci (quand le nombre de sites m tend vers l'infini).

Modèles à sites indépendants.

Dans le cas où les sites évoluent de façon indépendantes selon le même modèle, le calcul de la vraisemblance se ramène au calcul de la vraisemblance en chaque site, puisque dans ce cas :

$$L_m(\theta; x_{1:m}(T)) = \prod_{i=1}^{m} L_1(\theta, x_i(T)).$$

Dans le cas où l'évolution s'effectue de séquence à séquence, le calcul pour un site i de la vraisemblance $L_1(\theta; x_i(T))$ fait intervenir uniquement des exponentielles de matrices 4×4

d'après la forme des matrices de taux de sauts Q. Explicitement, on a :

$$L_1(\theta; x_i(T)) = \sum_{x_i(0)\in\mathcal{A}} P(x_i(0)) \times [\exp(TQ)](x_i(0), x_i(T)).$$

Dans le cas où l'évolution s'effectue sur un arbre, on note $G_{v\to v'}$ la matrice de transition de l'instant correspondant au nœud v à l'instant correspondant au nœud v', pour chaque arête orientée (v, v'). Pour tout nœud v, on note par $l(v)$ l'ensemble des nœuds feuilles de l'arbre issus de v.

En utilisant la propriété d'indépendance le long des arêtes de l'arbre on obtient pour chaque nœud v, pour chaque $x \in \mathcal{A}$ et $y \in \mathcal{A}^{\sharp l(v)}$:

– Si v est une feuille, $G_{v\to l(v)}(x, y) = \mathbf{1}(x = y)$.

– Sinon, on note v_1 et v_2 les deux nœuds fils de v et on pose :

$$G_{v\to l(v)}(x, y) = \left(\sum_{x_1\in\mathcal{A}} G_{v\to v_1}(x, x_1)G_{v_1\to l(v_1)}(x_1, y)\right)\left(\sum_{x_2\in\mathcal{A}} G_{v\to v_2}(x, x_2)G_{v_2\to l(v_2)}(x_2, y)\right).$$

On a alors, en notant r le nœud associé à la racine de l'arbre, que :

$$L_1(\theta; x_i(T)) = \sum_{x_i(0)\in\mathcal{A}} P(x_i(0)) \times G_{r\to l(r)}(x_i(0), x_i(T)).$$

Numériquement, l'algorithme de programmation dynamique de Felsenstein décrit dans [42] (section *Computing the Likelihood of a Tree*) permet de calculer efficacement la vraisemblance sur un arbre, grâce à une factorisation des termes à sommer.

En ce qui concerne le maximum de vraisemblance, le théorème de Wald fournit des conditions classiques d'existence, de consistance et de normalité asymptotique de celui-ci (voir le chapitre 5 de [114] et [116]).

Modèles à dépendance.

Pour un modèle avec dépendance aux voisins immédiats général et en particulier pour le modèle RN95+YpR, la situation est plus délicate car l'évolution de chaque site i dépend a priori de toute la séquence. En effet, par construction du modèle d'évolution, pour tout $\varepsilon > 0$ et dans la limite où ε tend vers 0, l'évolution du nucléotide au site i entre les instants $t-\varepsilon$ et t va dépendre des nucléotides aux sites $i-1$, i et $i+1$ à la date $t-\varepsilon$. On remarque alors que l'évolution du nucléotide en position $i-1$ (resp. $i+1$) entre les instants $t-2\varepsilon$ et $t-\varepsilon$ va dépendre quant à lui des nucléotides en positions $i-2$, $i-1$ et i (resp. en positions i, $i+1$ et $i+2$) à la date $t-2\varepsilon$. En répétant ce constat, on conclut que la dépendance peut se propager arbitrairement loin en général. On représente graphiquement ces dépendances sur la figure 2.1.

De ce fait, le calcul de la vraisemblance pour de tels modèles fait a priori appel au calcul d'une exponentielle de matrice de taille $4^m \times 4^m$, ce qui est impraticable numériquement sauf pour des séquences très courtes (de longueur plus petite que 8).

Dans ce cadre, un objectif principal est d'établir des méthodes permettant de fournir une approximation de la vraisemblance du modèle ou d'estimer le maximum de vraisemblance de la classe de modèles considérée. Deux approches possibles sont l'utilisation de méthodes par vraisemblances composites et l'utilisation de méthodes de type Monte-Carlo. Dans l'annexe C, des références complémentaires sont données sur les différentes approches

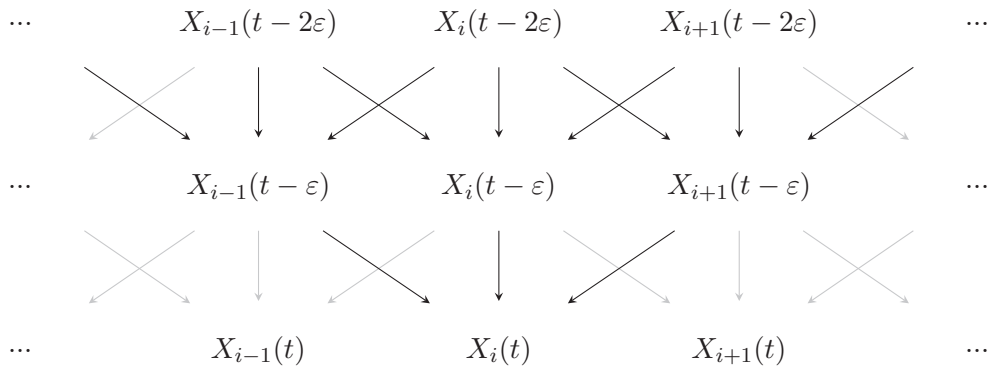


FIGURE 2.1 – Représentation schématique des dépendances dans un modèle avec dépendance aux voisins immédiats. Les arêtes en noir représentent les dépendances concernant le nucléotide $X_i(t)$.

permettant de prendre en compte les phénomènes de dépendance, dans le but de calculer ou d'approcher des quantités comme la vraisemblance.

Méthodes par vraisemblances composites. Les méthodes par vraisemblances composites (voir le chapitre 4 de cette thèse et le chapitre 2 de [47]) se séparent en deux parties principales : les vraisemblances composites conditionnelles et les vraisemblances composites marginales.

Les vraisemblances composites conditionnelles reposent sur une approximation de la structure de dépendance du modèle, en omettant certains termes. Pour l'étude des modèles d'évolution, des vraisemblances composites (ou des approximations de la vraisemblance dont la forme est proche de vraisemblances composites) ont été utilisées en négligeant la dépendance entre les sites qui ne sont pas voisins (voir [6, 7, 79, 105]). Le calcul de ces vraisemblances composites conditionnelles est rapide et permet d'obtenir également des estimations de maximums de vraisemblances composites associées. Par contre, l'écart entre l'approximation de la vraisemblance obtenue et la vraisemblance réelle n'est quantifiée qu'empiriquement.

Les vraisemblances composites marginales sélectionnent une certaine partie de l'information disponible, sans faire d'approximation. Les valeurs obtenues avec de telles vraisemblances composites ne sont plus comparables avec la vraisemblance réelle, mais des propriétés asymptotiques de consistance du maximum de vraisemblance composite associé sont connues. Dans [15], l'utilisation de ces vraisemblances permet d'obtenir des estimations consistantes et asymptotiquement normales du maximum de vraisemblance pour la classe de modèle RN95+YpR (cette approche est rappelée en détail dans la section 4.1).

Méthodes de type Monte-Carlo. D'autres approches pour calculer la vraisemblance du modèle et l'estimation du maximum de vraisemblance reposent sur des approximations de type Monte-Carlo, en traitant exactement le modèle original et ses dépendances. Elles se basent sur une structure de champ markovien d'ordre 2 du modèle (rappelée dans la

section 6.1.1). Dans [61] et [63], des échantillonnages de Gibbs puis une méthode d'inférence de type Monte-Carlo EM sont utilisées. Dans [9, 10, 11], l'estimation se fait avec une approche par intégration thermodynamique.

Ces méthodes fournissent des approximations de la vraisemblance convergentes ainsi que des estimations consistantes et asymptotiquement normales du maximum de vraisemblance. Par contre, le coût numérique de ces méthodes est important par rapport aux méthodes par vraisemblance composites.

Organisation des prochains chapitres.

Les propriétés spécifiques de RN95+YpR ne permettent pas de calculer explicitement la vraisemblance, mais rendent praticables certaines approches d'approximation de la vraisemblance.

Dans le chapitre 3, on présente des encodages déjà mis en évidence dans [14] permettant d'établir des propriétés pour les modèles RN95+YpR.

Dans le chapitre 4, on utilise ces encodages pour définir différentes vraisemblances composites associées aux modèles RN95+YpR. La vraisemblance composite marginale par triplets encodés est reprise de [15]. On introduit une vraisemblance composite conditionnelle appelée approximation markovienne. Ces vraisemblances composites sont étudiées théoriquement et on propose une estimation semi-empirique de l'écart d'estimation du maximum de vraisemblance dans le cas de la vraisemblance par triplets encodés.

Le chapitre 5 illustre les phénomènes de dépendance possibles de la classe RN95+YpR. On s'intéresse particulièrement aux phénomènes de dépendance extrêmes, i.e. lorsque certains renforcements de taux de sauts deviennent grands. Sur des exemples explicites, on montre en particulier que le calcul de la vraisemblance d'un nucléotide observé en un site peut dépendre de la valeur de tous les autres sites.

À l'aide des encodages spécifiques des modèles RN95+YpR, on développe dans le chapitre 6 de nouvelles structures markoviennes spatiales (cachées). En particulier, on montre une propriété de champ markovien d'ordre un (sur l'alphabet des nucléotides de cardinal 4) ainsi qu'une propriété de chaîne de Markov explicite d'ordre un sur un alphabet de cardinal 9.

Ces structures permettent d'abord de développer dans le chapitre 7 une preuve de convergence et de normalité asymptotique de l'estimateur du maximum de vraisemblance pour les modèles RN95+YpR. Elle repose sur un théorème générique issu de [21] pour les chaînes de Markov cachées, qui doit ici être adapté convenablement.

On cherche ensuite à utiliser ces structures pour approcher la vraisemblance d'un modèle RN95+YpR. La structure établie de chaîne de Markov d'ordre un rend possible l'utilisation de méthodes particulaires. Dans le chapitre 8, on décrit succinctement le principe des méthodes particulaires avant de détailler comment ces méthodes peuvent être mise en œuvre pour la structure étudiée.

L'implémentation effective en C++ d'une méthode particulaire est commentée dans le chapitre 9.

Enfin, on regroupe dans le chapitre 10 les applications numériques développées. On compare tout d'abord entre elles la consistance et la précision des trois approximations de vraisemblance étudiées : la vraisemblance composite par triplets encodés, la vraisemblance composite par approximation markovienne et l'approximation de la vraisemblance par méthodes particulaires. On compare ensuite les maximums de vraisemblance obtenus avec ces trois méthodes. On propose et teste également des méthodes d'inférence d'un nucléotide à la racine. Enfin, on effectue des comparaisons entre les modèles GTR, T92 et T92+CpGs sur deux jeux de séquences biologiques.

Chapitre 3

Encodages des séquences et modèles RN95+YpR

Dans ce chapitre, on présente différents types d'encodage des séquences d'ADN dont on cherche à décrire l'évolution. Ces encodages font apparaître des propriétés structurelles particulières des modèles RN95+YpR, mises en évidence dans [14], et dont l'importance est fondamentale pour la suite. Principalement, on montre que, après encodage, des portions de séquences disjointes évoluent de manière indépendante, selon une dynamique markovienne (voir proposition 3.2.7, théorème 3.2.10 et corollaire 3.5.2). Ce chapitre est essentiellement une présentation développée des propriétés déjà mises en évidence et étudiées dans [14].

L'organisation de ce chapitre est le suivant. Tout d'abord, on définit dans la section 3.1 l'encodage des nucléotides dans des alphabets restreints (encodages ρ, η et π) puis des séquences (séquences Φ-encodées ou π-encodées).

Ces encodages permettent de définir des évolutions π-encodées et Φ-encodées. Pour les modèles appartenant à la classe RN95+YpR, ces évolutions possèdent des propriétés spécifiques d'indépendance et leurs évolutions s'expriment à travers des matrices de taux de sauts instantanés sans dépendance ou avec dépendance vis-à-vis d'un seul voisin (section 3.2).

Lorsque l'évolution est conditionnée par une séquence finale qui n'est pas encodée, les propriétés énoncées ne sont plus valides et on explicite alors plusieurs contre-exemples illustratifs (section 3.3).

On remarque ensuite que les évolutions Φ-encodées donnent un moyen de gérer simplement les conditions aux extrémités de la séquence (section 3.4).

Enfin, on établit plusieurs conséquences importantes de l'utilisation d'encodages pour le calcul de la vraisemblance d'une séquence (section 3.5). On obtient notamment que l'on peut calculer explicitement la vraisemblance de séquences courtes Φ-encodées (de taille 2 à 6).Le cadre des évolutions Φ-encodées permet également d'écrire la vraisemblance d'une séquence Φ-encodée comme un produit de vraisemblance de séquences Φ-encodées minimales (corollaire 3.5.2). Ce dernier corollaire sera très utilisé dans les prochains chapitres, car il permet de ramener le calcul de la vraisemblance globale à celles de ces séquences Φ-encodées minimales.

Rappel des notations. On considère l'évolution d'une séquence de longueur m de séquence à séquence d'un instant 0 à un instant T, avec une séquence associée à la racine

fixée. On reprend la notation 1.2.1 utilisée pour décrire une telle évolution :

$$X = (X(t))_{t\in[0,T]} = (X_i)_{i\in[\![1,m]\!]} = (X_i(t))_{i\in[\![1,m]\!];t\in[0,T]}.$$

3.1 Encodages sur l'alphabet et les séquences

Encodages sur l'alphabet.

Définition 3.1.1. *On définit sur $\mathcal{A}$ les fonctions ρ qui confond les deux purines, η qui confond les deux pyrimidines et π qui identifie d'une part les deux purines et d'autre part les deux pyrimidines :*

$$\rho(A) := R\ ;\ \rho(G) := R\ ;\ \rho(C) := C\ ;\ \rho(T) := T, \tag{3.1}$$

$$\eta(A) := A\ ;\ \eta(G) := G\ ;\ \eta(C) := Y\ ;\ \eta(T) := Y, \tag{3.2}$$

$$\pi(A) := R\ ;\ \pi(G) := R\ ;\ \pi(C) := Y\ ;\ \pi(T) := Y. \tag{3.3}$$

Encodages sur les séquences.

Définition 3.1.2. *Pour une séquence $(x_1,\dots,x_m)\in\mathcal{A}^m$, on définit le Φ-encodage de cette séquence par :*

$$\Phi(x_1,\dots,x_m) = (\rho(x_1), x_2,\dots,x_{m-1},\eta(x_m)).$$

Notation 3.1.3. *L'espace d'états d'une séquence Φ-encodée est de cardinal $3\times 4^{m-2}\times 3$ et est donné par :*

$$S_m = \{R,C,T\}\times\mathcal{A}^{m-2}\times\{A,G,Y\}.$$

Remarque 3.1.4. *Une séquence Φ-encodée de longueur 3 est appelée triplet encodé. Il existe 36 triplets encodés différents.*

Remarque 3.1.5. *Le Φ-encodage pour les séquences de longueur 2 est appelé encodage (ρ,η). L'espace d'états est alors composé de 9 éléments notés (voir remarque 1.0.2 pour la notation CG vs CpG) :*

$$\mathcal{C} := \{RA, RG, RY, CA, CG, CY, TA, TG, TY\}.$$

3.2 Évolutions encodées et ambigües

3.2.1 Évolutions encodées

On regarde l'évolution d'une séquence de $\mathcal{A}^m$ à travers les encodages π, ρ et η. On fixe ici un modèle d'évolution M de type RN95+YpR. Tous les paramètres sont donc supposés fixés et on note $\theta = (v_x, w_x, r_y; x\in\mathcal{A}, y\in\mathcal{B})$ le jeu de paramètres associé (voir notation 1.2.4).

Définition 3.2.1. *Soit $i\in[\![1,m]\!]$. L'évolution π-encodée (resp. ρ-encodée, η-encodée) du site i correspond à l'évolution du site i, vue ensuite uniquement à travers la fonction π (resp. ρ, η).*

Cette évolution s'écrit $(\pi(X_i(t)))_{t\in[0,T]}$ (resp. $(\rho(X_i(t)))_{t\in[0,T]}$, $(\eta(X_i(t)))_{t\in[0,T]}$) avec la notation 1.2.1 de la séquence d'évolution.

Notation 3.2.2. *Dans le but d'alléger les notations, on pose pour tout site* $i \in [\![1, m]\!]$ *:*
- $\eta_i = \eta(X_i)$ *l'évolution* η*-encodée du site* i*.*
- $\rho_i = \rho(X_i)$ *l'évolution* ρ*-encodée du site* i*.*

On représente sur la figure 3.1 une évolution pendant une unité de temps d'un nucléotide, d'abord sans encodage puis avec les encodages π, ρ et η.

t	i
0	C
0.2	G
0.3	C
0.4	T
0.9	A
1	A

t	i
0	Y
0.2	R
0.3	Y
0.9	R
1	R

t	i
0	C
0.2	R
0.3	C
0.4	T
0.9	R
1	R

t	i
0	Y
0.2	G
0.3	Y
0.9	A
1	A

FIGURE 3.1 – Exemple d'évolution d'un nucléotide pendant une unité de temps, respectivement sans encodage, π-encodée, ρ-encodée et η-encodée.

3.2.2 Un cas particulier : l'encodage (ρ, η)

On remarque que la connaissance pour un site de η_i et de ρ_i permet de reconstituer entièrement l'évolution X_i. Ainsi, la connaissance de $(\rho_1, \eta_2, \rho_2, \ldots, \eta_{m-1}, \rho_{m-1}, \eta_m)$ permet d'obtenir l'ensemble de l'évolution Φ-encodée de $X_{1:m}$.

On définit l'évolution d'une séquence Φ-encodée de longueur 2 de la façon suivante :

Définition 3.2.3. *Pour chaque entier* $i \in [\![1, m-1]\!]$*, on appelle* Z_i *l'évolution* (ρ, η) *aux sites* $(i, i+1)$ *définie par :*

$$Z_i = (\rho_i, \eta_{i+1}).$$

On peut alors réécrire alors l'évolution Φ-encodée comme : $(Z_1, \ldots, Z_{m-1})$. La définition 3.2.3 sera souvent utilisée à partir du chapitre 4.

3.2.3 Évolutions ambigües

Pour comprendre les évolutions encodées, on souhaite étudier les matrices locales de taux de sauts associées. On définit alors trois processus de sauts, appelés évolution π-ambigüe (resp. ρ-ambigüe, η-ambigüe). On regarde ensuite les liens avec la famille locale de taux de sauts associée à l'évolution π-encodée (resp. ρ-encodée, η-encodée).

Définition 3.2.4. *Une évolution* π*-ambigüe associée au jeu de paramètres* θ *est un processus de sauts à deux états* $\{R, Y\}$ *et de matrice de taux de sauts instantanés :*

$$Q_\pi = \begin{array}{c} \\ R \\ Y \end{array} \begin{array}{c} \begin{array}{cc} R & Y \end{array} \\ \begin{pmatrix} . & v_C + v_T \\ v_A + v_G & . \end{pmatrix} \end{array}$$

où les termes diagonaux sont tels que la somme de chaque ligne est nulle.

Définition 3.2.5. *Une évolution ρ-ambigüe associée au jeu de paramètres θ et à la lettre $d \in \{A, G, Y\}$ est un processus de sauts à trois états $\{R, C, T\}$ et de matrice de taux de sauts instantanés :*

$$Q_{,d} = \begin{array}{c} \\ R \\ C \\ T \end{array} \begin{array}{c} \begin{array}{ccc} R & C & T \end{array} \\ \begin{pmatrix} . & v_C & v_T \\ v_A + v_G & . & \begin{array}{c} w_T + \\ r_{CA\to TA}\mathbf{1}_{d=A} + \\ r_{CG\to TG}\mathbf{1}_{d=G} \end{array} \\ v_A + v_G & \begin{array}{c} w_C + \\ r_{TA\to CA}\mathbf{1}_{d=A} + \\ r_{TG\to CG}\mathbf{1}_{d=G} \end{array} & . \end{pmatrix} \end{array}$$

où les termes diagonaux sont tels que la somme de chaque ligne est nulle.

Définition 3.2.6. *Une évolution η-ambigüe associée au jeu de paramètres θ et à la lettre $g \in \{R, C, T\}$ est un processus de sauts à trois états $\{A, G, Y\}$ et de matrice de taux de sauts instantanés :*

$$Q_{g,} = \begin{array}{c} \\ A \\ G \\ Y \end{array} \begin{array}{c} \begin{array}{ccc} A & G & Y \end{array} \\ \begin{pmatrix} . & \begin{array}{c} w_G + \\ r_{TA\to TG}\mathbf{1}_{g=T} + \\ r_{CA\to CG}\mathbf{1}_{g=C} \end{array} & v_T + v_C \\ \begin{array}{c} w_A + \\ r_{TG\to TA}\mathbf{1}_{g=T} + \\ r_{CG\to CA}\mathbf{1}_{g=C} \end{array} & . & v_T + v_C \\ v_A & v_G & . \end{pmatrix} \end{array}$$

où les termes diagonaux sont tels que la somme de chaque ligne est nulle.

Les processus de sauts des définitions 3.2.5 et 3.2.6 sont basés sur une combinaison entre :

- une matrice de taux de sauts à sites indépendants,
- le renforcement de certaines substitutions suivant la valeur de d (dans le cas de la définition 3.2.5) ou g (dans le cas de la définition 3.2.6).

3.2.4 Liens avec les évolutions encodées

Liens avec les évolutions π-encodées.

On montre maintenant que l'évolution π-encodée d'une séquence correspond exactement à choisir l'évolution π-ambigüe sur l'intervalle $[0, T]$ en chacun des sites.

Proposition 3.2.7. *L'évolution π-encodée d'une séquence associée au jeu de paramètres θ a la même loi que l'évolution suivante : pour $i \in [\![1, m]\!]$, chaque évolution au site i est régie de façon indépendante par une évolution π-ambigüe associée au jeu de paramètres θ.*

Démonstration. Les 8 dépendances aux voisins autorisées par le modèle RN95+YpR transforment toujours un dinucléotide YpR en un autre nucléotide YpR. Par exemple, la substitution CG vers TG n'est plus visible dans l'évolution π-encodée, s'écrivant dans les deux cas YR. Ainsi, l'évolution π-encodée n'est pas affectée par ces renforcements.

De plus, le modèle RN95 a comme propriété que le taux de saut d'une transversion ne dépend que du nucléotide obtenu après substitution. Par exemple, les taux de saut des transversions $A \to C$ et $G \to C$ sont égaux, ainsi que les taux de saut des transversions $A \to T$ et $G \to T$. Cette propriété permet d'établir les taux de sauts d'un nucléotide π-encodé vers un autre nucléotide π-encodé. On obtient alors la matrice correspondant à l'évolution π-ambigüe de la définition 3.2.4. □

Remarque 3.2.8. *On peut construire une preuve alternative en utilisant la construction de la dynamique par processus de Poisson indépendants (voir section 1.2.3), en montrant d'abord que pour tout site i, seuls les processus superposés : $\mathcal{V}_i^R := \mathcal{V}_i^A \cup \mathcal{V}_i^G$ et $\mathcal{V}_i^Y := \mathcal{V}_i^C \cup \mathcal{V}_i^T$ peuvent provoquer une transversion, et en concluant en remarquant que les mouvements de type V en un site ne dépendent pas de l'état des sites voisins.*

La proposition 3.2.7 et la remarque 3.2.8 permettent d'en déduire la proposition suivante, analogue à la construction de la section 1.2.3 :

Proposition 3.2.9. *Construction de l'évolution π-encodée par processus de Poisson.*

Pour chaque site i, on définit indépendamment :
- *$\mathcal{V}_i^R := \mathcal{V}_i^A \cup \mathcal{V}_i^G$ un processus homogène de Poisson sur $\mathbb{R}$ de taux $v_A + v_G$,*
- *$\mathcal{V}_i^Y := \mathcal{V}_i^C \cup \mathcal{V}_i^T$ un processus homogène de Poisson sur $\mathbb{R}$ de taux $v_C + v_T$.*

La séquence d'évolution π-encodée sur $[0, T]$ est notée $(\pi_1(t), \dots, \pi_m(t))_{t\in[0,T]}$, où m est la longueur de la séquence. À partir d'une séquence initiale $(\pi_1(0), \dots, \pi_m(0)) \in \{R, Y\}^m$, on définit l'évolution de la façon suivante.

Sur l'intervalle $[0, T]$, dès qu'une sonnerie d'un processus de Poisson associée à $\mathcal{V}_i^x$ se déclenche, le nucléotide au site i se substitue en x dans le cas où ce mouvement correspond à une transversion.

La dynamique ainsi définie coïncide avec celle décrite dans la proposition 3.2.7.

Liens avec les évolutions Φ-encodées.

Les évolutions ρ-ambigüe et η-ambigüe vérifient des énoncés analogues à la proposition 3.2.7. Nous allons surtout utiliser le théorème suivant.

Théorème 3.2.10. *Soit* M *un modèle inclus dans la classe de modèles RN95+YpR et θ son paramètre associé (voir notation 1.2.4).*

L'évolution Φ-encodée d'une séquence $(\rho(X_1)(t), X_2(t), \dots, X_{m-1}(t), \eta(X_m)(t))_{t\in[0,T]}$ issue du modèle M *a la même évolution que l'évolution markovienne dans le temps suivante : à chaque instant $t \in [0, T]$,*
- *le premier site suit la famille locale de taux de sauts ρ-ambigüe associée au jeu de paramètres θ et à la lettre $d = \eta(X_2(t))$, régie par la matrice de taux de sauts $Q_{,d}$,*
- *les sites intermédiaires suivent la matrice de taux de sauts $Q_{g,d}$ (voir définition 1.2.5),*

– *le dernier site suit la famille locale de taux de sauts η-ambigüe associée au jeu de paramètres θ et à la lettre $g = \rho(X_{m-1}(t))$, régie par la matrice de taux de sauts $Q_{g,}$.*

De plus :

– *l'évolution ne dépend pas des valeurs des nucléotides en dehors de ces sites (c'est-à-dire des sites non encodés strictement inférieurs à 1 ou strictement supérieurs à m),*

– *l'évolution ne dépend de la séquence initiale non encodée*

$$(X_1(0), X_2(0), \ldots, X_{m-1}(0), X_m(0))$$

qu'à travers

$$(\rho(X_1)(0), X_2(0), \ldots, X_{m-1}(0), \eta(X_m)(0)).$$

Le théorème 3.2.10 est prouvé dans [14].

3.3 Problèmes de conditionnement

On considère une évolution π-encodée (resp. ρ-encodée, η-encodée) conditionnée des séquences initiale et finale qui ne sont pas π-encodées (resp. ρ-encodées, η-encodées). Dans ce cas, la loi conditionnelle obtenue est en général différente de l'évolution π-ambigüe (resp. ρ-ambigüe, η-ambigüe) conditionnée par les mêmes séquences initiale et finale π-encodées (resp. ρ-encodées, η-encodées).

Donnons quelques contre-exemples illustratifs de cet énoncé.

Exemple 3.3.1. *Considérons $\varepsilon \ll 1$ et l'évolution d'un nucléotide en position notée i dans le modèle RN95 suivant (et donc sans dépendance aux voisins) :*

$$v_G = v_T = 1 \text{ et } w_x = v_{x'} = \varepsilon^2 \text{ pour } x \in \mathcal{A}, x' \in \{A, C\}.$$

On suppose que l'arête considérée a une longueur de ε et que le nucléotide initial est A, le nucléotide final est G. L'évolution va alors avoir avec grande probabilité le comportement suivant : deux substitutions vont se produire sur l'intervalle $[0, \varepsilon]$, la première substituant T à A et la seconde substituant G à T. On en déduit alors l'évolution π-encodée la plus probable.

Par contre, l'évolution π-ambigüe conditionnée par la même séquence finale π-encodée aboutit à effectuer l'évolution du nucléotide R vers le nucléotide R. Ainsi avec grande probabilité, aucune substitution ne va se produire.

Les deux évolutions sont donc différentes et cela est illustré par la figure 3.2, pour des instants $t_1, t_2 \in [0, \varepsilon]$.

Remarque 3.3.2. *En considérant le modèle de l'exemple 3.3.1 mais cette fois-ci avec le dinucléotide initial AC et le dinucléotide final GT, on obtient la même différence de comportement, et les nucléotides initiaux et finaux encodés dans π sont égaux à RY.*

Exemple 3.3.3. *On reprend le modèle de l'exemple 3.3.1 en ajoutant le renforcement suivant :*

$$r_{TG \to TA} = 1/\varepsilon.$$

On choisit encore une arête de longueur de ε. On considère le dinucléotide Φ-encodé initial RG et le dinucléotide Φ-encodé final RA. On remarque que le dinucléotide final considéré encodé par les fonctions π et ρ est dans les deux cas RR, et que $RR \neq RA$.

t	i
0	A
t_1	T
t_2	G
ε	G

t	i
0	R
t_1	Y
t_2	R
ε	R

t	i
0	R
ε	R

FIGURE 3.2 – Comportement le plus probable pour l'exemple 3.3.1 de respectivement l'évolution conditionnée par le nucléotide G, l'évolution π-encodée conditionnée par le nucléotide G et l'évolution π-ambigüe conditionnée par le nucléotide $\pi(G) = R$ (cette dernière évolution est aussi l'évolution π-encodée conditionnée par le nucléotide R).

On regarde comment l'évolution (dans l'encodage Φ) va se comporter sur l'intervalle $[0, \varepsilon]$. D'après le théorème 3.2.10, on peut regarder uniquement les évolutions ambigües. Le nucléotide de droite veut aller de G vers A. Or, cela n'est probable que lorsque le nucléotide à sa gauche est T. Ainsi, comme les transversions vers T et les transversions vers R sont probables, avec grande probabilité l'évolution va subir trois substitutions (dans l'encodage Φ).

La première substituant T à R sur le premier nucléotide, la seconde substituant A à G sur le deuxième nucléotide et la troisième substituant R à T sur le premier nucléotide. On en déduit alors les évolutions π-encodée et ρ-encodée les plus probables.

Par contre, l'évolution π-ambigüe (resp. ρ-ambigüe) conditionnée par la même séquence finale π-encodée (resp. ρ-encodée) aboutit à effectuer l'évolution du nucléotide R vers le nucléotide R sur les deux nucléotides. Ainsi avec grande probabilité, aucune substitution ne va se produire.

Les deux évolutions sont donc encore différentes et cela est illustré par les figures 3.3 et 3.4.

t	i	i+1
0	R	G
t_1	T	
t_2		A
t_3	R	
ε	R	A

t	i	i+1
0	R	R
t_1	Y	
t_3	R	
ε	R	R

t	i	i+1
0	R	R
ε	R	R

FIGURE 3.3 – Comportement le plus probable pour l'exemple 3.3.3 de respectivement l'évolution conditionnée par le dinucléotide RA, l'évolution π-encodée conditionnée par le dinucléotide RA et l'évolution π-ambigüe conditionnée par le dinucléotide $\pi(RA) = RR$ (cette dernière évolution est aussi l'évolution π-encodée conditionnée par le dinucléotide RR).

t	i	i+1
0	R	G
t_1	T	
t_2		A
t_3	R	
ε	R	A

t	i	i+1
0	R	R
t_1	T	
t_3	R	
ε	R	R

t	i	i+1
0	R	R
ε	R	R

FIGURE 3.4 – Comportement le plus probable pour l'exemple 3.3.3 de respectivement l'évolution conditionnée par le dinucléotide RA, l'évolution ρ-encodée conditionnée par le dinucléotide RA et l'évolution ρ-ambigüe conditionnée par le dinucléotide $\rho(RA) = RR$ (cette dernière évolution est aussi l'évolution ρ-encodée conditionnée par le dinucléotide RR).

3.4 Précision sur la gestion de la condition aux bords

Pour gérer les conditions aux bords avec la condition I ou la condition II (voir la section 1.2.2), la façon la plus simple et que l'on utilise le plus souvent est de considérer que les séquences traitées sont déjà Φ-encodées, et qu'elles ne sont pas connues de façon plus précise. Dans ce cas, le théorème 3.2.10 montre que les conditions aux bords I et II sont équivalentes.

Remarque 3.4.1. *Dans le cas où les nucléotides initiaux et finaux de la séquence associé au temps final sont dans respectivement $\{C, T\}$ et $\{A, G\}$, les conditions aux bords I et II sont encore équivalentes puisque les séquences Φ-encodées sont alors égales aux séquences non encodées.*

Si on veut éviter de perdre l'information sur les bords de la séquence introduite en Φ-encodant les séquences, on utilise la condition aux bords II sur la séquence initiale. Dans ce cas, la remarque 3.4.1 montre que si on considère l'évolution Φ-encodée sur les sites $[\![0, m+1]\!]$, avec les feuilles non fixées aux sites 0 et $m+1$, l'évolution induite sur $[\![1, m]\!]$ coïncide avec l'évolution non encodée avec la condition aux bords II.

3.5 Conséquences pour le calcul de la vraisemblance

Le théorème 3.2.10 établi dans le cadre des évolutions Φ-encodées permet de déduire l'expression de la vraisemblance d'une séquence Φ-encodée.

Corollaire 3.5.1. *Calcul exact de la vraisemblance d'une séquence Φ-encodée. On rappelle que l'on considère l'évolution de séquence à séquence sur l'intervalle $[0, T]$ et que le nombre de sites est donné par $m \geq 2$. Pour un modèle M inclus dans la classe de modèles RN95+YpR, on note $\tilde{Q}$ la matrice de taux de sauts associée à l'évolution décrite dans le théorème 3.2.10 des séquences Φ-encodées de longueur m.*

La vraisemblance d'obtenir la séquence observée Φ-encodée :

$$\Phi(x(T)) = (\rho(x_1)(T), x_2(T), \ldots, x_{m-1}(T), \eta(x_m)(T)) \in \{C, T, R\} \times \mathcal{A}^{m-2} \times \{A, G, Y\}$$

sachant que la séquence à l'instant initial est :

$$\Phi(x(0)) = (\rho(x_1)(0), x_2(0), \ldots, x_{m-1}(0), \eta(x_m)(0))$$

est alors donnée par :

$$L(\Phi(x(T))) = (\exp(T\tilde{Q}))(\Phi(x(0)), \Phi(x(T))).$$

On va voir que cette expression peut être calculée numériquement sur des séquences courtes (section 3.5.1) et peut être découpée en produit de morceaux minimaux (section 3.5.2).

3.5.1 Calcul pour les séquences courtes encodées

Le corollaire 3.5.1 permet de calculer numériquement la vraisemblance des observations de séquences Φ-encodées, à travers le calcul d'exponentielles de matrices. Néanmoins, le calcul effectif n'est possible que pour les séquences de longueurs courtes puisque la taille de $\tilde{Q}$ croît exponentiellement en fonction du nombre de sites m ($\tilde{Q}$ est de taille $9.4^{m-2} \times 9.4^{m-2}$).

Notons que :
- pour $m = 2$, $\tilde{Q}$ est une matrice 9×9,
- pour $m = 3$, $\tilde{Q}$ est une matrice 36×36.

3.5.2 Découpage en produits minimaux

La remarque 3.4.1 et le théorème 3.2.10 permettent de montrer que la vraisemblance d'une séquence peut s'écrire comme le produit de vraisemblance de séquences Φ-encodées minimales (en terme de nombre de sites).

Corollaire 3.5.2. *Découpage RY. Soit $i \in [\![1, m-1]\!]$. On suppose que le modèle est fixé et que la séquence associée à la racine est fixée.*

La vraisemblance d'une séquence Φ-encodée $(\rho(y_1), y_2 \ldots, y_{m-1}, \eta(y_m)) \in S_m$ vérifiant $\rho(y_i) = R$ et $\rho(y_{i+1}) = Y$ est égale au produit des vraisemblances des séquences Φ-encodées $(\rho(y_1), y_2, \ldots, y_{i-1}, \eta(y_i))$ et $(\rho(y_{i+1}), y_{i+2}, \ldots, y_{m-1}, \eta(y_m))$.

Démonstration. On note L la fonction de vraisemblance associée au modèle et à la racine. L'évolution est décrite par les variables $(\rho(X_1), X_2, \ldots, X_{m-1}, \eta(X_m))$ avec :

$$(\rho(X_1), X_2, \ldots, X_{m-1}, \eta(X_m))(T) = (\rho(y_1), y_2 \ldots, y_{m-1}, \eta(y_m)).$$

Comme $\rho(y_i) = R$ et $\rho(y_{i+1}) = Y$, on a $\eta(y_i) = y_i$ et $\rho(y_{i+1}) = y_{i+1}$. On peut alors écrire :

$$L(\rho(y_1), y_2 \ldots, y_{m-1}, \eta(y_m)) = L(\rho(y_1), y_2, \ldots, y_{i-1}, \eta(y_i), \rho(y_{i+1}), y_{i+2}, \ldots, y_{m-1}, \eta(y_m)).$$

Or, par le théorème 3.2.10, les évolutions de $(X_1, \ldots, \eta(X_i))$ et $(\rho(X_{i+1}), \ldots, X_m)$ sont indépendantes. Ainsi, la vraisemblance ce décompose en la forme énoncée dans le corollaire :

$$L(\rho(y_1), y_2 \ldots, y_{m-1}, \eta(y_m)) = L(\rho(y_1), y_2 \ldots, y_{i-1}, \eta(y_i)) \times L(\rho(y_{i+1}), y_{i+2} \ldots, \eta(y_m)).$$

□

Exemple 3.5.3. *Si la séquence observée est $CCTAGCTATCCGTA$, le calcul de la vraisemblance de cette séquence se décompose comme le produit des vraisemblances des séquences $CCTAG$, CTA, $TCCG$ et TA.*

Remarque 3.5.4. *Lorsque la séquence à la racine n'est pas fixée, la propriété de découpage RY des vraisemblances n'est pas nécessairement vérifiée. Néanmoins, si on choisit la loi stationnaire du modèle à la racine ou un modèle à sites indépendants, le corollaire 3.5.2 reste valide. En effet si on choisit la loi stationnaire du modèle à la racine, le calcul de la vraisemblance du corollaire 3.5.2 ne nécessite que la connaissance de la loi stationnaire de l'évolution $(\rho(X_1), X_2, \ldots, X_{i-1}, \eta(X_i), \rho(X_{i+1}), X_{i+2}, \ldots, X_{m-1}, \eta(X_m))$, qui se décompose en deux morceaux par le théorème 3.2.10.*

Le corollaire 3.5.2 de découpage d'une séquence en morceaux indépendants est à rapprocher avec la vraisemblance composite par découpage de données – ou *split data likelihood* [100] – consistant à découper la séquence en morceaux de taille fixée puis en approchant la vraisemblance comme le produit des vraisemblances de chaque morceau. Le fait important du corollaire 3.5.2 est qu'il donne, pour des morceaux dont la longueur dépend des observations, une méthode de découpage de la séquence qui n'altère pas le calcul de la vraisemblance complète.

Chapitre 4

Vraisemblances composites

Dans ce chapitre, on s'intéresse à divers types de vraisemblances composites pour les modèles inclus dans la classe RN95+YpR. Il s'agit de quantités qui ne correspondent pas à la véritable vraisemblance associée au couple (modèle, observations), mais dont on s'attend à ce qu'elles se comportent de manière similaire, tout en étant effectivement calculables. Ces quantités sont utilisées pour contourner les difficultés numériques du calcul de la vraisemblance exacte de séquences (voir le chapitre 2).

Les vraisemblances composites sont définies de la façon suivante :

Définition 4.0.5. *Pour un vecteur aléatoire $(Y_i)_{i\in[\![1,m]\!]}$ et un ensemble de paramètres Θ, on note L la fonction de vraisemblance. On choisit des événements marginaux ou conditionnels $A_1, \dots, A_K$ liés au vecteur $(Y_i)_{i\in[\![1,m]\!]}$ et on associe pour $k \in [\![1, K]\!]$ sa vraisemblance L_k. Explicitement, on pose pour $\theta \in \Theta$ et $y_{1:m} \in (Y_i)_{i\in[\![1,m]\!]}$:*

$$L_k(\theta; y_{1:m}) = L(\theta; y_{1:m} \in A_k).$$

Une vraisemblance composite L_C est alors définie, pour des poids $(w_k)_{k\in[\![1,K]\!]}$ positifs, par :

$$L_C(\theta; y_{1:m}) = \prod_{k=1}^{K} L_k(\theta; y_{1:m})^{w_k}.$$

Pour obtenir une valeur calculable numériquement, chaque événement marginal ou conditionnel est choisi pour simplifier le calcul global de la vraisemblance composite. En pratique, on distingue les vraisemblances composites marginales (composées uniquement d'événements marginaux) et les vraisemblances composites conditionnelles (composées uniquement d'événements conditionnels).

Exemple 4.0.6. *La vraisemblance composite par découpage de données avec chevauchement – ou* split data likelihood with overlapping blocks *[47, 100] – est une vraisemblance composite marginale consistant à découper la séquence en morceaux de taille fixée se chevauchant. Pour des morceaux de longueur 2, on retrouve une vraisemblance composite par paire [74], et elle est alors définie par :*

$$L_C^{ex1}(\theta; y_{1:m}) = \prod_{i=1}^{m-1} P_\theta(y_i, y_{i+1}).$$

La vraisemblance composite d'ordre k [8] est une vraisemblance composite conditionnelle qui suppose que la dépendance d'un site vis-à-vis des précédents est markovienne d'ordre k. Pour $k = 1$, elle est donnée par :

$$L_C^{ex2}(\theta; y_{1:m}) = \prod_{i=1}^{m-1} P_\theta(y_{i+1}|y_i).$$

Les vraisemblances composites ont été introduites par Lindsay dans [76], mais leur utilisation est déjà présente dans [16] sous le nom de pseudo-vraisemblance. Une vue d'ensemble du sujet a été réalisée dans [115] en 2011. Elles ne sont en général pas comparables directement avec la vraisemblance complète, mais leur maximum de vraisemblance l'est parfois : dans deux cas particuliers, [4] et [47] fournissent des propriétés asymptotiques du maximum de vraisemblance composite. Une étude des vraisemblances composites dans le cadre des chaînes de Markov cachées est réalisée dans [47].

Les propriétés spécifiques des modèles RN95+YpR liées aux encodages (en particulier la section 3.5) permettent un calcul effectif de diverses vraisemblances composites. Dans ce chapitre, on étudie principalement deux types de vraisemblances composites, basées sur les couples $(Z_i(T))_{i\in[\![1,m-1]\!]}$ de nucléotides observés Φ-encodés (voir définition 3.2.3).

- La vraisemblance composite par triplets encodés, introduite dans [15] et implémenté dans le logiciel `bppml` [15] de Bio++ [38, 39]. Elle correspond à la vraisemblance composite marginale par paires associée aux couples $(Z_i(T))_{i\in[\![1,m-1]\!]}$. La valeur obtenue n'est pas directement comparable avec la vraisemblance exacte du modèle. Par contre, des propriétés spécifiques du modèle permettent de prouver la consistance et la normalité asymptotique du maximum de vraisemblance composite associé, ainsi que d'obtenir une expression théorique de la variance asymptotique.
- Les vraisemblances composites par approximation markovienne d'ordre k. Elles correspondent aux vraisemblances composites conditionnelles d'ordre k associées aux couples $(Z_i(T))_{i\in[\![1,m-1]\!]}$. Elles fournissent des approximations de la vraisemblance directement comparables avec la vraisemblance réelle du modèle.

Ce chapitre est séparé en deux parties. Dans la section 4.1, on décrit la vraisemblance composite par triplets encodés et les propriétés asymptotiques vérifiées par cette vraisemblance. Ensuite, on établit une expression théorique de la variance asymptotique du maximum de vraisemblance composite par triplets encodés. On fournit également une méthode numérique semi-empirique permettant d'utiliser effectivement cette expression dans une estimation numérique de type Monte-Carlo.

Dans la section 4.2, on décrit les vraisemblances composites par approximation markovienne d'ordre k et on montre une propriété asymptotique de consistance locale des maximums de vraisemblance associés.

Utilisation des vraisemblances composites.

Aspects numériques. L'utilisation des vraisemblances composites par triplets encodés et par approximations markoviennes pour approcher la vraisemblance exacte et pour estimer les paramètres du modèle est détaillée dans le chapitre 10 consacré aux applications.

Identifiabilité de la classe de modèle RN95+YpR. Les propriétés d'indépendance entre chaque séquence d'une phase de la vraisemblance composite par triplets encodés permettent de fournir des conditions d'identifiabilité des modèles RN95+YpR. La définition de l'identifiabilité et les résultats obtenus sont placés dans l'annexe B.

Hypothèses et notations utilisées dans ce chapitre. On choisit Θ inclus dans l'ensemble des paramètres d'évolutions possibles du modèle RN95+YpR (inclus dans $\mathbb{R}^{16}$).

Hypothèses.

Hypothèse 4.0.7. *On suppose fixé un modèle RN95+YpR paramétré par $\theta_0 \in \Theta$ et l'ensemble des séquences associées aux feuilles de l'arbre. On suppose que l'ensemble Θ des paramètres est compact. De plus, on suppose que θ_0 est dans l'intérieur de Θ.*

Concernant la loi à la racine, on utilise l'hypothèse suivante.

Hypothèse 4.0.8. *On étudie les cas où la loi à la racine R_0 est une séquence fixée, la loi stationnaire du modèle M ou une approximation de la loi stationnaire de M.*

Dans les deux premiers cas, la loi à la racine vérifie la propriété d'indépendance des marginales Φ-encodées du théorème 3.2.10, c'est-à-dire qu'en tout site i et pour tout k, la variable aléatoire $(\rho_i(0), \dots, \eta_{i+k}(0))$ est indépendante des variables $(\rho_1(0), \dots, \eta_{i-1}(0))$ et $(\rho_{i+k+1}(0), \dots, \eta_m(0))$. De plus, dans les deux premiers cas, la probabilité d'une séquence à la racine est également infiniment différentiable (lorsque l'on consdière la loi stationnaire du modèle, la probabilité s'exprime comme limite uniforme en θ d'exponentielles de matrices).

Dans le cas d'une approximation de la loi stationnaire de M, on suppose que les propriétés d'indépendance des marginales Φ-encodées et de dérivibabilité sont vérifiées.

La topologie de l'arbre et les différentes longueurs de branches T_0 sont fixées.

Notations. On considère la suite $(Z_i)_{i\in[\![1,m-1]\!]} = (\rho_i, \eta_{i+1})_{i\in[\![1,m-1]\!]}$ (voir définition 3.2.3) issue du modèle et on utilise les notations suivantes, avec T l'instant final.

Notation 4.0.9. *Pour $l \geq 1$ fixé, le $(l+1)$-uplet Φ-encodé au temps final associé aux sites $(i, \dots, i+l)$ est noté :*

$$W_{i,l} := (\rho_i, X_{i+1}, \dots, X_{i+l-1}, \eta_{i+l})(T) = (Z_i(T), \dots, Z_{i+l-1}(T)).$$

Pour $(w_{i,l})_{i\in[\![1,m-l]\!]}$ une réalisation de $(W_{i,l})_{i\in[\![1,m-l]\!]}$, la probabilité de $w_{i,l}$ sous un modèle $\theta \in \Theta$ est notée :

$$\iota_{\theta,l}(w_{i,l}) := P_\theta(W_{i,l} = w_{i,l}).$$

Dans le cas particulier des triplets, on allège les notations par : $W_{i,2} = W_i$ et pour $(w_i)_{i\in[\![1,m-2]\!]}$ une réalisation de $(W_i)_{i\in[\![1,m-2]\!]}$, $\theta \in \Theta$, on note :

$$\iota_{\theta,2}(w_i) = \iota_\theta(w_i).$$

Remarque 4.0.10. *Comme la loi à la racine vérifie la propriété d'indépendance des marginales Φ-encodées du théorème 3.2.10, les vraisemblances $\iota_{\theta,l}(w_{i,l})$ peuvent être exprimées et manipulées explicitement :*

$$\iota_{\theta,l}(w_{i,l}) = \sum_r P(W_{i,l}(0) = r) P(W_{i,l} = w_{i,l} | W_{i,l}(0) = r),$$

et le terme $P(W_{i,l} = w_{i,l} | W_{i,l}(0) = r)$ s'exprime sur un arbre comme une somme d'exponentielles de matrices (voir aussi le paragraphe Calcul de chaque terme *de la section 4.1.2).*

4.1 Triplets encodés

4.1.1 Construction et propriétés asymptotiques

La méthode suivante est issue de [15] (paragraphe *averaged log-likelihood*). On ne fait ici aucune approximation par rapport au modèle RN95+YpR initial, dans le sens où on considère toujours des évolutions issues d'un modèle RN95+YpR pour lequel on calcule la vraisemblance de marginales de ces évolutions. La vraisemblance composite considérée se base sur le découpage triplets encodés par triplets encodés sans chevauchement de la séquence. Il y a trois manières de découper la séquence en triplets encodés, suivant la position $p \in \{1, 2, 3\}$ (appelée phase) du premier triplet considéré. Pour une séquence de longueur 11, on représente ces trois phases sur la figure 4.1.

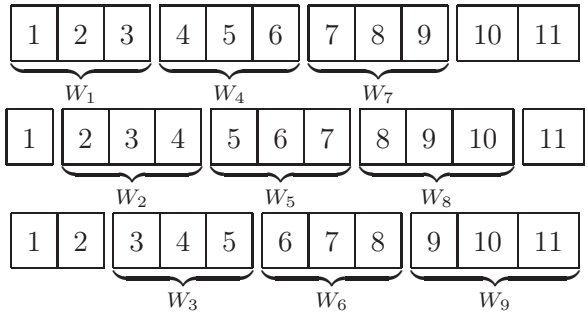


FIGURE 4.1 – Les triplets encodés considérés pour les phases respectivement 1, 2 et 3, d'une séquence de longueur 11.

Pour chaque phase, par le théorème 3.2.10, les triplets considérés sont indépendants et on définit les vraisemblances composites suivantes, pour les phases $p \in \{1, 2, 3\}$:

$$\begin{aligned}\hat{L}^p(\theta) &= P\left((Z_{3i+p}(T), Z_{3i+p+1}(T))_{i\in[\![0,\lfloor(m-2-p)/3\rfloor]\!]}\right)\\ &= \prod_{i=0}^{\lfloor(m-2-p)/3\rfloor} P(Z_{3i+p}(T), Z_{3i+p+1}(T))\\ &= \prod_{i=0}^{\lfloor(m-2-p)/3\rfloor} \iota_\theta(W_{3i+p}),\end{aligned}$$

On définit enfin :

Définition 4.1.1. *La vraisemblance composite par triplets encodés est :*

$$\hat{L}_{triplets}(\theta) = \hat{L}^1(\theta)\hat{L}^2(\theta)\hat{L}^3(\theta).$$

Remarque 4.1.2. *On peut réécrire :* $\hat{L}_{triplets}(\theta) = \prod_{i=1}^{m-2} \iota_\theta(W_i)$. *La quantité calculée est donc une vraisemblance composite marginale.*

La proposition suivante permet de valider l'utilisation de cette vraisemblance composite (proposition montrée dans [15], paragraphe *A theoretical justification*).

Proposition 4.1.3. *On se place sous l'hypothèse 4.0.7. On suppose que tous les paramètres $\theta \in \Theta$ sont identifiables pour le modèle d'une des phases de triplets encodés (c'est-à-dire, avec les notations de l'annexe B, de l'injectivité de $\lambda \mapsto P_{\infty,\lambda}^{triplets}$, avec $\lambda \in (\mathsf{R}_0, \mathsf{T}_0, \mathsf{M}_\theta)_{\theta\in\Theta}$).*

La vraisemblance composite moyenne $\frac{1}{m}\hat{L}_{triplets}(\theta)$ fournit un estimateur consistant et asymptotiquement normal (quand m tend vers l'infini) de l'espérance de la vraisemblance d'un trinucléotide Φ-encodé.

De plus, l'estimateur associé de maximum de vraisemblance composite estime de façon consistante le paramètre θ lorsque la longueur de la séquence tend vers l'infini.

Démonstration. Idée de la preuve. On considère chacune des phases $\hat{L}^1(\theta)$, $\hat{L}^2(\theta)$ et $\hat{L}^3(\theta)$. Comme chaque vraisemblance composite $\hat{L}^p(\theta)$ (pour $p \in \{1,2,3\}$) s'exprime comme un produit de termes issues de variables aléatoire (W_{3i+p}) indépendantes et identiquement distribuées (par les propriétés d'indépendance du théorème 3.2.10), on en déduit la consistance et la normalité asymptotique de $\frac{1}{m}\hat{L}^p(\theta)$ vers l'espérance de la vraisemblance d'un trinucléotide Φ-encodé.

On en déduit ensuite les mêmes propriétés pour la moyenne de ces trois estimateurs par des propriétés de mélange du modèle et l'hypothèse de stationnarité spatiale de la distribution des observations. □

Remarque 4.1.4. ***Cas des l-uplets.*** *Dans le cas où on ne considère non pas des triplets mais des l-uplets (pour $l \geq 2$), on définit de façon analogue la vraisemblance composite par l-uplets encodés :*

$$\hat{L}_{l\text{-}uplets}(\theta) = \prod_{i=1}^{m-l+1} \iota_{\theta,l-1}(W_{i,l-1}).$$

La proposition 4.1.3 reste valide.

La vraisemblance composite par l-uplets encodés est un cas particulier de vraisemblance composite par paires définie dans [74]. En effet, en identifiant $W_{i,l}$ avec $(W_{i,l-1}, W_{i+1,l-1})$ pour $i \in [\![1, m-l]\!]$ et $l \geq 2$ (avec en particulier $W_{i,2} = (Z_i(T), Z_{i+1}(T))$), la vraisemblance composite par l-uplets encodés correspond à la vraisemblance composite par paires avec décalage $L = 1$.

4.1.2 Variance asymptotique du maximum de vraisemblance composite

Dans cette section, on précise le comportement de l'estimateur de maximum de vraisemblance composite triplets par triplets. Pour cela, on exprime l'écart asymptotique entre le paramètre estimé $\hat{\theta}_0^m$ et le paramètre à rechercher θ_0. On fournit ensuite une méthode semi-empirique de calcul de l'écart-type de l'estimateur. Dans la section 10.5.3, cette méthode sera utilisée et comparée à l'écart-type empirique direct sur deux modèles RN95+YpR.

On rappelle que l'on effectue l'évolution le long de m sites notés $[\![1, m]\!]$. On considère des paramètres d'évolution du modèle RN95+YpR, variant dans un compact $\Theta \subset \mathbb{R}^{16}$. On suppose de nouveau que ces paramètres sont identifiables pour le modèle d'une des phases de triplets encodés. La topologie de l'arbre et les différentes longueurs de branches sont fixées. On note $\mathsf{M}(\theta)$ le modèle associé au paramètre θ.

On fixe un paramètre θ_0 qui n'est pas sur le bord de Θ et on suppose que l'évolution est effectuée selon le modèle $\mathsf{M}(\theta_0)$. On obtient alors des séquences associées aux feuilles

regroupées triplets par triplets $(w_i)_{i\in[\![1,m-2]\!]}$ (voir notation 4.0.9 avec $l = 2$) issues de $(W_i)_{i\in[\![1,m-2]\!]}$. On note aussi W une variable aléatoire de loi ι_{θ_0}.

Remarque 4.1.5. *Dans le cadre fixé, la fonction $\theta \mapsto \iota_\theta$ est infiniment différentiable. En effet, conditionnellement à la connaissance de la racine, l'évolution s'exprime sous forme d'exponentielles de matrices dont les coefficients sont des combinaisons linéaires issues θ. La racine est également infiniment différentiable d'après l'hypothèse 4.0.8.*

Notation 4.1.6. *Pour tout $\theta \in \Theta$, on utilise les notations suivantes :*
- $l_m(\theta) := (m-2)E_{\theta_0} \log \iota_\theta(W)$,
- $\hat{l}_m(\theta) := \log \hat{L}_{triplets}(\theta) = \sum_{i=1}^{m-2} \log \iota_\theta(W_i)$,
- $\hat{\theta}_0^m = \text{argmax}_{\theta\in\Theta}\, \hat{l}_m(\theta)$,
- $H(\hat{l}_m)(\hat{\theta}_0^m)$ *la matrice hessienne associée à* $\hat{l}_m$ *en* $\hat{\theta}_0^m$.
- *Pour une matrice symétrique réelle positive M, on note $\sqrt{M}$ la matrice symétrique réelle positive vérifiant $(\sqrt{M})^2 = M$.*

Expression de l'écart asymptotique de $\hat{\theta}_0^m - \theta_0$. La log-vraisemblance composite triplets par triplets des observations $(W_i)_{i\in[\![1,m-2]\!]}$ sous le modèle $\mathsf{M}(\theta)$ s'exprime par $\hat{l}_m(\theta)$. On estime θ_0 par l'estimateur du maximum de vraisemblance composite triplets par triplets $\hat{\theta}_0^m$ et la proposition 4.1.3 montre que :

$$\hat{\theta}_0^m \to \theta_0.$$

Le théorème suivant permet de préciser le comportement de l'écart $\hat{\theta}_0^m - \theta_0$.

Théorème 4.1.7. *On pose pour tout $\theta \in \Theta$ et tout i : $f_\theta = \frac{\nabla \iota_\theta}{\iota_\theta}$, $U_i = f_\theta(W_i) - E_{\theta_0} f_\theta(W)$, les matrices de variance-covariance suivantes (pour tout m) :*

$$v_{\theta,\theta_0}^m = \frac{1}{m-2} E_{\theta_0}\left[\left(\sum_{i=1}^{m-2} U_i\right)\left(\sum_{i=1}^{m-2} U_i\right)^T\right],$$

et enfin :

$$v_{\theta,\theta_0} = \lim_{m\to+\infty} v_{\theta,\theta_0}^m.$$

L'existence de cette dernière quantité résulte de l'équation (4.4).

Soit ξ un vecteur aléatoire réel normal centré réduit. Quand le nombre de sites m tend vers l'infini, la convergence en loi suivante est vérifiée :

$$\frac{1}{\sqrt{m-2}}[H(\hat{l}_m)(\hat{\theta}_0^m)](\hat{\theta}_0^m - \theta_0) \to \sqrt{v_{\theta_0,\theta_0}}\xi.$$

Calcul explicite de chaque terme dans le cas $d = 1$. Avant de montrer le théorème, on souhaite expliciter chacun des termes présents dans le théorème 4.1.7 dans le cas où un seul paramètre d'évolution du modèle est à rechercher. On considère que ce paramètre varie dans un intervalle $\Theta \subset \mathbb{R}_*^+$ pour lequel il y a identifiabilité, les valeurs de tous les autres paramètres étant supposées connues. Chacun des termes décrit peut être calculé numériquement et deux exemples issus de ces calculs sont proposés dans la section 10.5.3.

– Les quantités ι_θ, ι'_θ et ι''_θ peuvent être calculées explicitement pour tout $\theta \in \Theta$. En effet, pour une racine fixée, ι_θ s'exprime comme une exponentielle de matrice 36×36 et $\iota'_\theta, \iota''_\theta$ comme leurs dérivées. Ainsi le calcul de ces quantités se déduit du calcul d'exponentielles de matrices de taille 108×108 (voir [48]).
– La dérivée seconde $\hat{l}''_m(\theta)$ s'exprime ensuite par :

$$\hat{l}''_m(\theta) = \sum_{i=1}^{m-2} \frac{\iota''_\theta(W_i)}{\iota_\theta(W_i)} - \sum_{i=1}^{m-2} \left(\frac{\iota'_\theta(W_i)}{\iota_\theta(W_i)} \right)^2 .$$

– Enfin, en utilisant la portée de dépendance d'au plus deux pas de $(W_i)_{i\in[\![1,m-2]\!]}$, la variance v^m_{θ,θ_0} vérifie :

$$\begin{aligned}
(m-2)v^m_{\theta,\theta_0} =& E_{\theta_0}\left[\left(\sum_{i=1}^{m-2} U_i\right)^2\right] && (4.1)\\
=& E_{\theta_0}\left[\sum_{i=1}^{m-2} U_i^2 + 2\sum_{i=1}^{m-2}\sum_{i'=i+1}^{m-2} U_i U_{i'}\right] && (4.2)\\
=& E_{\theta_0}\left[\sum_{i=1}^{m-2} U_i^2 + 2\sum_{i=1}^{m-4}(U_iU_{i+1} + U_iU_{i+2}) + 2U_{m-3}U_{m-2}\right] && (4.3)\\
=& (m-2)E_{\theta_0}[U_1^2] + 2(m-3)E_{\theta_0}[U_1U_2] + 2(m-4)E_{\theta_0}[U_1U_3]. && (4.4)
\end{aligned}$$

On estime $E_{\theta_0}[U_1^2]$ (resp. les différents termes $E_{\theta_0}[U_iU_{i'}]$) par la variance empirique (resp. les covariances empiriques) obtenue à partir des données $(w_i)_{i\in[\![1,m-2]\!]}$. Le calcul de l'estimation de la variance est semi-empirique dans le sens où seule la quantité v_{θ_0,θ_0} est estimée empiriquement.

Remarque 4.1.8. ***Calcul de ι_θ, ι'_θ et ι''_θ pour un arbre.*** *Sur un arbre, on considère l'ensemble $S = S_3$ des triplets encodés (voir aussi notation 3.1.3). Pour tout nœud v, on note par $l(v)$ l'ensemble des nœuds feuilles de l'arbre issus de v.*

On définit par récurrence la matrice de transition sur l'arbre de la même manière que lors du calcul de la vraisemblance pour un modèle à sites indépendants (voir chapitre 2). Pour cela, on note $G_{v\to v'}$ la matrice de transition de l'instant correspondant au nœud v à l'instant correspondant au nœud v', pour chaque arête orientée (v, v'). Pour tout nœud v, on note par $l(v)$ l'ensemble des nœuds feuilles de l'arbre issus de v. Pour chaque nœud v, $x \in S$ et $y \in S^{\sharp l(v)}$ l'ensemble des séquences observées aux feuilles, on a :

– *Si v est une feuille, $G_{v\to l(v)}(x, y) = \mathbf{1}(x = y)$.*
– *Sinon, on note v_1 et v_2 les deux nœuds fils de v et on pose :*

$$G_{v\to l(v)}(x,y) = \left(\sum_{x_1\in\mathcal{A}} G_{v\to v_1}(x,x_1)G_{v_1\to l(v_1)}(x_1,y)\right)\left(\sum_{x_2\in\mathcal{A}} G_{v\to v_2}(x,x_2)G_{v_2\to l(v_2)}(x_2,y)\right).$$

On a alors, en notant r le nœud associé à la racine de l'arbre et $w_i(T)$ l'ensemble de triplets observés aux feuilles de l'arbre, que :

$$\iota_\theta(w_i(T)) = \sum_{w_i(0)\in S} P_\theta(w_i(0))G_{r\to l(r)}(w_i(0), w_i(T)).$$

Pour obtenir ι'_θ et ι''_θ, on exprime les dérivées $G'_{v\to l(v)}$ et $G''_{v\to l(v)}$ par rapport au paramètre θ. Par abus de notation on omet de mentionner les indices des fonctions G, G' et G''. Pour chaque nœud v, $x \in S$ et $y \in S^{\sharp l(v)}$, on a :

– *Si v est une feuille, $G'_{v\to l(v)}(x,y)=0$ et $G''_{v\to l(v)}(x,y)=0$.*
– *Sinon, on a :*

$$G'_{v\to l(v)}(x,y)=\left(\sum_{x_1\in S}G'(x,x_1)G(x_1,y)+G(x,x_1)G'(x_1,y)\right)\left(\sum_{x_2\in S}G(x,x_2)G(x_2,y)\right)$$
$$+\left(\sum_{x_1\in S}G(x,x_1)G(x_1,y)\right)\left(\sum_{x_2\in S}G'(x,x_2)G(x_2,y)+G(x,x_2)G'(x_2,y)\right),$$

$$G''_{v\to l(v)}(x,y)=$$
$$\left(\sum_{x_1\in S}G(x,x_1)G(x_1,y)\right)\left(\sum_{x_2\in S}G''(x,x_2)G(x_2,y)+2G'(x,x_2)G'(x_2,y)+G(x,x_2)G''(x_2,y)\right)$$
$$+2\left(\sum_{x_2\in S}G'(x,x_2)G(x_2,y)+G(x,x_2)G'(x_2,y)\right)\left(\sum_{x_1\in S}G'(x,x_1)G(x_1,y)+G(x,x_1)G'(x_1,y)\right)$$
$$+\left(\sum_{x_2\in S}G(x,x_2)G(x_2,y)\right)\left(\sum_{x_1\in S}G''(x,x_1)G(x_1,y)+2G'(x,x_1)G'(x_1,y)+G(x,x_1)G''(x_1,y)\right).$$

On en déduit ensuite ι'_θ et ι''_θ.

Démonstration du théorème 4.1.7. Pour des suites de variables aléatoires réelles R_m, X_m et Y_m telles que $Y_m\to 0$ (resp. Y_m bornée) en probabilité et $X_m=Y_mR_m$, on écrit :

$$X_m=o(R_m)\text{ (resp. }X_m=O(R_m)).$$

Démonstration. On a θ_0 appartenant à l'intérieur de Θ. Ainsi pour tout m, on a d'une part que :

$$\nabla l_m(\theta_0)=0$$

(par interversion entre l'opérateur de dérivation et l'intégrale) et d'autre part que :

$$\nabla\hat{l}_m(\hat{\theta}_0^m)=0.$$

(par définition de $\hat{\theta}_0^m$). On écrit alors :

$$\nabla\hat{l}_m(\hat{\theta}_0^m)-\nabla\hat{l}_m(\theta_0)+\nabla\hat{l}_m(\theta_0)-\nabla l_m(\theta_0)=0. \tag{4.5}$$

On réécrit les deux premiers termes de (4.5) en faisant intervenir la matrice hessienne :

$$\nabla\hat{l}_m(\hat{\theta}_0^m)-\nabla\hat{l}_m(\theta_0)=[H(\hat{l}_m)(\hat{\theta}_0^m)](\hat{\theta}_0^m-\theta_0)+(m-2)o(\hat{\theta}_0^m-\theta_0).$$

Or, $\hat{\theta}_0^m-\theta_0=O\left(\frac{1}{\sqrt{m}}\right)$ (cf section 5.3 de [114] par exemple). On écrit donc :

$$\nabla\hat{l}_m(\hat{\theta}_0^m)-\nabla\hat{l}_m(\theta_0)=[H(\hat{l}_m)(\hat{\theta}_0^m)](\hat{\theta}_0^m-\theta_0)+o(\sqrt{m}).$$

Pour les deux derniers termes de (4.5), on va identifier un théorème central limite pour la variable $\nabla \hat{l}_m(\theta_0)$, d'une manière similaire à la proposition 4.3 de [40].

On observe que $(f_\theta(W_i))_i$ est mélangeante (puisque pour tout i, W_i est indépendante de $(W_{i+k})_{k\geq 3}$). Cela implique l'ergodicité de cette suite (voir par exemple le chapitre 6.4 de [37]). En réutilisant que W_i est indépendante de $(W_{i+k})_{k\geq 3}$, on vérifie les conditions du théorème 5.2 de Hall et Heyde [56] et on obtient que v^m_{θ,θ_0} converge vers une matrice positive et le théorème central limite suivant :

$$\frac{1}{\sqrt{m-2}}\left(\sum_{i=1}^{m-2} f_\theta(W_i) - (m-2)E_{\theta_0} f_\theta(W)\right) \to \sqrt{v_{\theta,\theta_0}}\xi$$

ce qui correspond à :

$$\frac{1}{\sqrt{m-2}}\left(\nabla \hat{l}_m(\theta) - \nabla l_m(\theta)\right) \to \sqrt{v_{\theta,\theta_0}}\xi.$$

En regroupant dans (4.5), on obtient le résultat souhaité :

$$\frac{1}{\sqrt{m-2}}[H(\hat{l}_m)(\hat{\theta}^m_0)](\hat{\theta}^m_0 - \theta_0) \to \sqrt{v_{\theta_0,\theta_0}}\xi.$$

□

4.2 Approximations markoviennes

Dans la section 4.1, on a considéré des vraisemblances composites marginales du modèle RN95+YpR pour lesquelles on a calculé la vraisemblance. Ces quantités ne sont pas des approximations de la vraisemblance du modèle RN95+YpR global. On considère dans cette section des vraisemblances composites fournissant des approximations de la vraisemblance du modèle RN95+YpR. On verra que ces quantités sont reliées simplement aux vraisemblances composites obtenues par l-uplets encodés (voir remarque 4.1.4), ce qui fournira des propriétés asymptotiques de consistance locale.

L'approximation markovienne à $l \in \mathbb{N}^*$ pas (cas typique $l = 1$) cherche à approcher la vraisemblance d'une séquence :

$$P(z_{1:m-1}(T)) = P(z_1(T)) \prod_{i=2}^{m-1} P(z_i(T)|z_{1:i-1}(T))$$

par une chaîne de Markov à l pas. Pour cela, on remplace dans le calcul de la vraisemblance chaque terme $P(z_i(T)|z_{1:i-1}(T))$ par la quantité $P(z_i(T)|z_{i-l:i-1}(T))$ que l'on est capable de calculer. On note $(\tilde{Z}_i)_{i\in[\![1,m-1]\!]}$ la chaîne de Markov à l pas associée. Par le théorème 3.2.10, on sait que la loi de $(\tilde{Z}_i, \dots \tilde{Z}_{i+l})$ est égale à la loi de $(Z_i(T), \dots Z_{i+l}(T))$.

On explicite dans la proposition suivante le comportement de la chaîne $(\tilde{Z}_i)_{i\in[\![1,m-1]\!]}$ en fonction des quantités utilisées pour définir la vraisemblance par triplets encodés.

Notation 4.2.1.

- *On rappelle que d'après les notations 4.0.9,* $W_{i,l} = (Z_i(T), \dots, Z_{i+l-1}(T))$.
- *On pose pour tous* i *et* l *:* $\tilde{w}_{i,l} = (\tilde{z}_i, \dots, \tilde{z}_{i+l-1})$.

– *Pour simplifier la preuve, on écrit $\tilde{z}_i$ à la place de $\tilde{Z}_i = \tilde{z}_i$ pour $i \in [\![1, m-1]\!]$.*

Proposition 4.2.2. *Pour $\theta \in \Theta$ et $(\tilde{z}_j)_{j\in[\![1,m-1]\!]}$ réalisation de $(\tilde{Z}_j(T))_{j\in[\![1,m-1]\!]}$ chaîne de Markov à l pas, on a :*

$$P_\theta(\tilde{z}_j|\tilde{z}_{j-1},\dots,\tilde{z}_1) = \begin{cases} \iota_{\theta,1}(\tilde{w}_{1,1}) & \text{si } j=1, \\ \iota_{\theta,j}(\tilde{w}_{1,j})/\iota_{\theta,j-1}(\tilde{w}_{1,j-1}) & \text{si } j \in [\![2,l]\!], \\ \iota_{\theta,l+1}(\tilde{w}_{j-l,l+1})/\iota_{\theta,l}(\tilde{w}_{j-l,l}) & \text{si } j \in [\![l+1,m-1]\!]. \end{cases}$$

Cela implique :

$$P_\theta(\tilde{z}_1,\dots,\tilde{z}_{m-1}) = \iota_{\theta,l}(\tilde{w}_{1,l})\frac{\hat{L}_{l+2\text{-}uplets}(\theta)}{\hat{L}_{l+1\text{-}uplets}(\theta)}\iota_{\theta,l}(\tilde{w}_{m-l,l}).$$

Démonstration.

$$\begin{aligned} P(\tilde{z}_1,\dots,\tilde{z}_{m-1}) =& P(\tilde{z}_1)P(\tilde{z}_2 \mid \tilde{z}_1)P(\tilde{z}_3 \mid \tilde{z}_2,\tilde{z}_1)\dots P(\tilde{z}_{m-1} \mid \tilde{z}_{m-2},\dots,\tilde{z}_1) \\ =& P(\tilde{z}_1)P(\tilde{z}_2 \mid \tilde{z}_1)P(\tilde{z}_3 \mid \tilde{z}_2,\tilde{z}_1)\dots P(\tilde{z}_{m-1} \mid \tilde{z}_{m-2},\dots,\tilde{z}_{m-l-1}) \\ =& P(\tilde{z}_1)\prod_{j=1}^{l-1} P(\tilde{z}_{j+1}|\tilde{z}_j,\dots,\tilde{z}_1)\prod_{i=1}^{m-l-1} P(\tilde{z}_{l+i}|\tilde{z}_{l+i-1},\dots,\tilde{z}_i) \\ =& P(W_{1,1}=\tilde{w}_{1,1})\prod_{j=1}^{l-1}\frac{P(W_{1,j+1}=\tilde{w}_{1,j+1})}{P(W_{1,j}=\tilde{w}_{1,j})}\prod_{i=1}^{m-l-1}\frac{P(W_{i,l+1}=\tilde{w}_{i,l+1})}{P(W_{i,l}=\tilde{w}_{i,l})} \\ =& P(W_{1,l}=\tilde{w}_{1,l})\frac{\prod_{i=1}^{m-l-1}P(W_{i,l+1}=\tilde{w}_{i,l+1})}{\prod_{i=1}^{m-l}P(W_{i,l}=\tilde{w}_{i,l})}P(W_{m-l,l}=\tilde{w}_{m-l,l}) \\ =& P(W_{1,l}=\tilde{w}_{1,l})\frac{\hat{L}_{l+2\text{-uplets}}(\theta)}{\hat{L}_{l+1\text{-uplets}}(\theta)}P(W_{m-l,l}=\tilde{w}_{m-l,l}). \end{aligned}$$

Dans la deuxième égalité, on a utilisé la définition des $(\tilde{z}_j)_{j\in[\![1,m-1]\!]}$. Dans la quatrième égalité, on a utilisé que la loi de $(\tilde{Z}_i,\dots\tilde{Z}_{i+l})$ est égale à la loi de $(Z_i(T),\dots Z_{i+l}(T))$. Dans la dernière égalité, on a utilisé la définition des vraisemblances composites obtenus par l-uplets encodés (voir remarque 4.1.4). □

Définition 4.2.3. *On appelle $\hat{L}_{l\text{-}Markov}$ vraisemblance composite par approximation markovienne à l pas, pour $l \geq 1$. Pour le cas particulier $l=1$, on écrit simplement $\hat{L}_{Markov}$.*

Remarque 4.2.4. *En particulier, pour $l=1$, la vraisemblance de $(\tilde{z}_j)_{j\in[\![1,m-1]\!]}$ vérifie :*

$$P_\theta(\tilde{z}_j|\tilde{z}_{j-1},\dots,\tilde{z}_1) = \begin{cases} \iota_{\theta,1}(\tilde{w}_{1,1}) & \text{si } j=1, \\ \iota_{\theta,2}(\tilde{w}_{j-1,2})/\iota_{\theta,1}(\tilde{w}_{j-1,1}) & \text{si } j \in [\![2,m-1]\!]. \end{cases}$$

et :

$$\hat{L}_{Markov}(\theta) = P_\theta(\tilde{z}_1,\dots,\tilde{z}_{m-1}) = \iota_{\theta,1}(\tilde{w}_{1,1})\frac{\hat{L}_{triplets}(\theta)}{\hat{L}_{couples}(\theta)}\iota_{\theta,1}(\tilde{w}_{m-1,1}).$$

Remarque 4.2.5. *On ne considère pas le cas $l=0$ car on veut conserver pour tout i la structure de dépendance entre Z_i et Z_{i+1}, qui existe même pour un modèle d'évolution i.i.d. (cf la construction à cheval des $Z_i = (\rho_i, \eta_{i+1})$ de la définition 3.2.3). Une autre approximation de la vraisemblance est mentionnée à la fin de cette section.*

Remarque 4.2.6. *Écart générique entre la vraisemblance composite par approximation markovienne et la vraisemblance exacte. Pour $k \geq 1$, on considère un modèle d'évolution RN95+YpR avec dépendance et une séquence d'observations $y_{1:k+3}$ de longueur $k+3$ débutant par le nucléotide π-encodé R, terminant par Y, et tel que l'écart d'estimation entre la log-vraisemblance estimée par approximation markovienne et la log-vraisemblance exacte V soit non nulle. L'existence d'une telle séquence d'observation n'est pas démontrée de façon générale, mais est vérifiée sur les modèles de l'annexe A pour $k \in \{1, 2, 3\}$. On note alors $r > 0$ cet écart de log-vraisemblance.*

Pour $l \geq 1$, on choisit un nombre de sites $m = l(k+3)$ et on construit la séquence observée par l répétitions de la séquence $y_{1:k+3}$. On obtient alors par le corollaire 3.5.2 de découpage de la vraisemblance en morceaux indépendants que l'écart de vraisemblance vaut $\frac{r}{k+3}m$.

L'erreur commise par l'estimation par approximation markovienne pour ces observations est donc de l'ordre de m et la proportion d'écart de log-vraisemblance $\frac{L-\hat{L}_{k\text{-}Markov}}{L}$ vaut $\frac{r}{(k+3)V} \neq 0$.

Proposition 4.2.7. *$\theta \mapsto \hat{L}_{Markov}(\theta)$ admet un extremum local asymptotique (en la longueur m de la séquence) en θ_0.*

Démonstration. $\theta \mapsto \frac{\partial}{\partial\theta} \log \hat{L}_{\text{triplets}}(\theta)$ et $\theta \mapsto \frac{\partial}{\partial\theta} \log \hat{L}_{\text{couples}}(\theta)$ s'annulent en θ_0 quand la longueur de la séquence m tend vers l'infini, donc $\theta \mapsto \frac{\partial}{\partial\theta} \log \hat{L}_{\text{Markov}}(\theta)$ également. □

Autre approximation de la vraisemblance. On a vu que la construction à cheval des $Z_i = (\rho_i, \eta_{i+1})$ suggère de conserver une structure de dépendance entre les variables composites $\tilde{Z}_i$ et $\tilde{Z}_{i+1}$. Les approximations précédentes se basaient sur des $(l+2)$-uplets avec $l \geq 1$. On va voir ici que l'on peut définir une vraisemblance composite gardant les mêmes propriétés que les vraisemblances composites par approximations markoviennes et basée sur les couples encodés de dinucléotides.

L'approximation markovienne à $\frac{1}{2}$ pas consiste à remplacer dans le calcul de la vraisemblance chaque terme $P(z_i(T)|z_{1:i-1}(T))$ par la quantité $P(z_i(T)|\pi_i(T))$. On note encore $(\tilde{Z}_i)_{i\in[\![1,m-1]\!]}$ la chaîne de Markov associée.

Soit $\theta \in \Theta$ et $(\tilde{z}_j)_{j\in[\![1,m-1]\!]}$ réalisation de $(\tilde{Z}_j(T))_{j\in[\![1,m-1]\!]}$.

Définition 4.2.8. *On pose $\hat{L}_{pi}(\theta)$ la vraisemblance du modèle π-encodée, c'est-à-dire la quantité :*

$$\prod_{i=1}^{m} P(\pi_i(T)).$$

La vraisemblance de $(\tilde{z}_j)_{j\in[\![1,m-1]\!]}$ vérifie alors la proposition suivante :

Proposition 4.2.9. *On a :*

$$P_\theta(\tilde{z}_1, \ldots, \tilde{z}_{m-1}) = P(\pi_1(T))\frac{\hat{L}_{couples}(\theta)}{\hat{L}_{pi}(\theta)}.$$

Démonstration. Preuve similaire à la proposition 4.2.2. □

Taille des matrices manipulées. Pour l'utilisation pratique de la vraisemblance composite par approximation markovienne à l pas, la table 4.1 récapitule la taille des matrices manipulées (pour lesquelles on doit calculer les exponentielles sur les différentes branches de l'arbre).

l	taille des matrices manipulées	
1/2	9×9	et 2×2
1	36×36	et 9×9
2	144×144	et 36×36
3	576×576	et 144×144
4	2304×2304	et 576×576

TABLE 4.1 – Taille des matrices à considérer pour calculer une vraisemblance composite par approximation markovienne à l pas, pour différentes valeurs de l.

Chapitre 5

Dépendance le long d'une séquence

Dans ce chapitre, on cherche à comprendre les différents phénomènes de dépendance présents le long d'une séquence issue d'un modèle RN95+YpR.

Commençons par récapituler les résultats théoriques positifs obtenus dans le chapitre 3 en terme de dépendance le long de la séquence observée $X_{1:m}(T)$. D'après le théorème 3.2.10, on a la propriété d'indépendance suivante :

$$P(X_{i:m}(T)|X_{1:i-2}(T)) = P(X_{i:m}(T)).$$

De plus, cette propriété peut être renforcée dans le cas particulier où $(\rho_{i-1}, \eta_i) = RY$ d'après le corollaire 3.5.2, et on obtient alors :

$$P(X_{i:m}(T)|X_{1:i-1}(T)) = P(X_{i:m}(T)) \quad \text{si} \quad (X_{i-1}(T), X_i(T)) \in \{AC, GC, AT, GT\}.$$

Au vu des propriétés de l'encodage, on peut facilement en déduire des conclusions erronées. En choisissant des modèles particuliers, on établit dans ce chapitre les résultats négatifs suivants, qui illustrent des phénomènes de dépendance extrêmes à portée longue :

- le comportement des séquences observées $(X_i(T))_{i\in[\![1,m]\!]}$ n'est pas markovien en général, à n'importe quel ordre fixé (section 5.1.2),
- la valeur obtenue par approximation markovienne (à n'importe quel ordre fixé) peut même fournir des valeurs de vraisemblance non pertinentes, dans le sens où la proportion d'écart de log-vraisemblance entre la vraisemblance exacte et l'approximation markovienne $\frac{L-\hat{L}_{k\text{-Markov}}}{L}$ tend vers 100% quand le nombre de sites tend vers l'infini (section 5.1.3),
- l'inférence d'un nucléotide à la racine peut dépendre de l'ensemble de la séquence observée (section 5.1.4),
- la loi stationnaire du modèle n'est pas markovienne en général (section 5.2).

Pour montrer ces résultats, on choisit des modèles définis comme des situations limites de modèles RN95+YpR pour lesquels il est possible d'effectuer des calculs théoriques exacts, principalement les modèles des exemples 5.1.1 et 5.2.2.

Notons que l'étude de la dépendance le long d'une séquence sera complétée dans le chapitre 10 consacré aux applications numériques, sur des modèles typiques ou atypiques non limites.

5.1 Étude d'un modèle d'évolution limite

Dans toute cette section, on considère les 3 modèles suivants (qui ne diffèrent que par leurs lois à la racine) pour lesquels on va identifier différents phénomènes de dépendance.

Exemple 5.1.1. *On choisit le modèle d'évolution RN95+YpR donné par*

$$v_G = 1,\ r_{CG\to CA} \to +\infty,$$

où les coefficients non indiqués sont nuls. L'évolution est considérée de séquence à séquence, sur une longueur temporelle infinie et une longueur spatiale m. La racine est quant à elle choisie parmi les 3 choix suivants :

- *la racine est fixée par :*

$$C\underbrace{C\dots C}_{m-1}, \tag{5.1}$$

- *la racine est fixée par :*

$$G\underbrace{C\dots C}_{m-1}, \tag{5.2}$$

- *la loi à la racine est la loi uniforme parmi (5.1) et (5.2).*

5.1.1 Propriétés du modèle

L'évolution des modèles de l'exemple 5.1.1 sont décrits de la façon suivante :

Propriété 5.1.2. *Pour un site i et une date t_0 fixés :*

- *Si le nucléotide est égal à A, alors pour toutes dates $t > t_0$ le nucléotide reste égal à A.*
- *Si le nucléotide est égal à G, alors :*
  - *si le nucléotide $i-1$ au temps t_0 est C, alors le nucléotide au site i se substitue en A à cet instant.*
  - *si le nucléotide $i-1$ au temps t_0 n'est pas C, alors pour toutes dates $t > t_0$ le nucléotide $i-1$ ne sera jamais égal à C. Ainsi pour toutes dates $t > t_0$ le nucléotide reste égal à G.*
- *Si le nucléotide est égal à C ou T, alors le nucléotide se substitue en G avec un taux de substitution* 1.

La propriété 5.1.2 permet de lier chaque évolution à racine fixée à un ensemble de permutations et d'instants de changements. On énonce la propriété pour le cas particulier de la séquence associée à la racine (5.1). Dans la remarque 5.1.6, on obtient un résultat analogue avec la racine (5.2).

Propriété 5.1.3. *Depuis la séquence associée à la racine (5.1), l'évolution est entièrement déterminée par l'ordre des substitutions $C \to G$ et les instants de substitution.*

Connaître l'évolution correspond alors à se donner :

- *d'une part une permutation de longueur m où à chaque site on associe son rang de substitution ;*
- *d'autre part un ensemble d'instants de changements $t_1 < t_2 < \ldots < t_m$.*

Exemple 5.1.4. *Depuis la séquence associée à la racine (5.1), on considère les instants de changements* $t_1 < \ldots < t_5$ *et la permutation* $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 1 & 5 & 4 \end{pmatrix}$.

On reconstitue alors l'évolution donnée sur la figure 5.1.

t	1	2	3	4	5
0	C	C	C	C	C
t_1			G		
t_1+			A		
t_2	G				
t_3		G			
t_4					G
t_4+					A
t_5				G	
$+\infty$	G	G	A	G	A

FIGURE 5.1 – Évolution associée à l'exemple 5.1.4.

La séquence finale est entièrement déterminée par l'étude des variations de la permutation associée comme le montre la propriété suivante :

Propriété 5.1.5. *On note* σ *la permutation associée à l'évolution depuis la séquence associée à la racine (5.1). On a au temps final :*

- *le nucléotide au site* 1 *est* G.
- *Pour* $i \in [\![2, m]\!]$, *le nucléotide au site* i *est* G *si et seulement si* $\sigma(i) > \sigma(i-1)$, *et est* A *sinon.*

Démonstration. Soit $i \in [\![2, m]\!]$. Pour avoir G en position i, il faut qu'à l'instant de changement le nucléotide au site $i-1$ ne soit pas égal à C. Il faut ainsi que la substitution vers G ait déjà eu lieu en position $i-1$. Donc $\sigma(i) > \sigma(i-1)$. □

Remarque 5.1.6. *Si on part de la séquence associée à la racine (5.2), on observe que cela correspond à fixer* $\sigma(1) = 1$ *partant de la racine (5.1). Ainsi, on associe à chaque évolution une permutation de* $\mathfrak{S}_{m-1}$.

5.1.2 Probabilité d'un nucléotide sachant les observations passées

Dans cette section, on exhibe un modèle pour lequel l'évolution des séquences observées n'est pas markovienne d'ordre k (pour $k \in \mathbb{N}^*$). Pour cela, on considère de nouveau l'exemple 5.1.1 avec une séquence de longueur $m = a+b$, et la séquence associée à la racine donnée par (5.1) :

$$\underbrace{C \ldots C}_{a+b}.$$

La séquence observée pour les $a+b-1$ premiers sites est donnée par :

$$G\underbrace{G\dots G}_{a-1}\underbrace{A\dots A}_{b-1} \tag{5.3}$$

ou

$$G\underbrace{A\dots A}_{a-1}\underbrace{A\dots A}_{b-1} \tag{5.4}$$

On établit que la probabilité pour que le nucléotide au site final m soit égal à A dépend du choix de la séquence observée (5.3) ou (5.4), quelle que soit la valeur de b. De plus, on explicite l'écart de probabilité entre les deux probabilités obtenues.

Propriété 5.1.7. *La probabilité que le nucléotide feuille du site $a+b$ soit A est :*
- *$\frac{a+b-1}{(a+b)b}$ si la séquence observée des premiers sites est (5.3).*
- *$\frac{1}{a+b}$ si la séquence observée des premiers sites est (5.4).*

Démonstration. On utilise la propriété 5.1.5. Pour le calcul de $P(G\underbrace{G\dots G}_{a-1}\underbrace{A\dots A}_{b})$, on cherche les permutations $\sigma \in \mathfrak{S}_{a+b}$ avec $a-1$ montées suivies de b descentes. Cela impose $\sigma(a) = a+b$ donc on dénombre $\binom{a+b-1}{a-1}$ permutations possibles. On a alors : $P(G\underbrace{G\dots G}_{a-1}\underbrace{A\dots A}_{b}) = \frac{1}{(a+b)(a-1)!b!}$ et la probabilité recherchée vaut :

$$\frac{a+b-1}{(a+b)b}.$$

Pour la deuxième probabilité, on regarde les permutations $\sigma \in \mathfrak{S}_{a+b}$ décroissantes $a+b-1$ fois. Il n'y a qu'un seul choix possible et la probabilité recherchée est :

$$\frac{1}{a+b}.$$

□

Conséquence 5.1.8. *L'écart entre les deux probabilités de la propriété 5.1.7 vaut :*

$$\Delta = \frac{a-1}{(a+b)b}.$$

En particulier, quand $a \to +\infty$, on a : $\Delta \sim 1/b$.
Quand $a = 2$ et $b \to +\infty$, on a : $\Delta \sim 1/b^2$.

5.1.3 Approximations markoviennes et valeur exacte

On cherche à mettre en défaut l'estimation de la vraisemblance par approximation markovienne en utilisant le modèle de l'exemple 5.1.1. Pour cela, on calcule pour des observations fixées et pour toute longueur de séquence m les valeurs de log-vraisemblances exactes et les approximations markoviennes à k pas associées.

On montre que l'erreur commise sur la log-vraisemblance en utilisant la méthode par approximation markovienne à k pas fixé peut être de l'ordre de $m \log m$ quand le nombre de sites m tend vers l'infini. On en déduit que la proportion d'écart de log-vraisemblance entre la vraisemblance exacte et l'approximation markovienne $\frac{L-\hat{L}_{k\text{-Markov}}}{L}$ tend vers 100% quand le nombre de sites tend vers l'infini.

Remarque 5.1.9. *Cette valeur est à comparer avec l'erreur d'estimation générique de la vraisemblance composite par approximation markovienne par rapport à la vraisemblance exacte.*

D'après la remarque 4.2.6, on sait que pour des modèles d'évolution avec dépendance, on peut construire des observations telles que la proportion d'écart de log-vraisemblance entre la valeur exacte et l'estimation par approximation markovienne à k pas soit une constante non nulle quand le nombre de sites m tend vers l'infini.

Exemple 5.1.10. *On choisit le modèle de l'exemple 5.1.1 avec (5.1) comme séquence à la racine (constituée uniquement de nucléotides C) et de la séquence observée suivante :*

$$G\underbrace{A\dots A}_{m-1}.$$

Par la propriété 5.1.7, on sait que pour $i \geq 1$, la probabilité d'obtenir $R\underbrace{A\dots A}_{i-1}$ sachant que les $i-1$ premiers nucléotides sont $R\underbrace{A\dots A}_{i-2}$ est de $1/i$.

On peut en déduire les valeurs exactes de la vraisemblance et des estimations par approximations markoviennes, énoncées dans la propriété suivante :

Propriété 5.1.11. *Pour le modèle et l'observation de l'exemple 5.1.10 :*
- *la vraisemblance de la séquence est donnée par $1/m!$,*
- *l'estimation par approximation markovienne à $1/2$ pas est donnée par $1/2^{m-1}$,*
- *l'estimation par approximation markovienne à k pas est donnée pour $k \geq 1$ par :*

$$1/\left((k+1)!(k+2)^{m-1-k}\right).$$

En conséquence, à un nombre de pas k fixé, l'écart entre la la log-vraisemblance exacte et l'estimation par approximation markovienne est équivalent à $m \log m$ et la proportion d'écart et la proportion d'écart $\frac{L-\hat{L}_{k\text{-Markov}}}{L}$ converge vers 1 quand le nombre de sites m tend vers l'infini. L'estimation par approximation markovienne n'est alors plus pertinente pour estimer la log-vraisemblance. On représente sur la figure 5.2 les log-vraisemblances obtenues pour des séquences de longueurs 10 et 100, avec $k \in \{1/2, 1, 2, 3, 4\}$ ainsi que la log-vraisemblance exacte.

5.1.4 Inférence à la racine

Dans cette section, on cherche à inférer un nucléotide de la séquence ancestrale à partir d'un modèle et d'une séquence d'observations. Sur deux exemples, on montre que l'inférence de ce nucléotide dépend de l'ensemble de la séquence d'observations, et on explicite les inférences obtenues en fonction de la longueur de la séquence d'observations.

Premier exemple.

On considère le modèle de l'exemple 5.1.1 avec comme loi à la racine la loi uniforme parmi les deux séquences (5.1) et (5.2), et les deux séquences associées aux feuilles suivantes :

$$\underbrace{G\dots G}_{m-1}G, \tag{5.5}$$

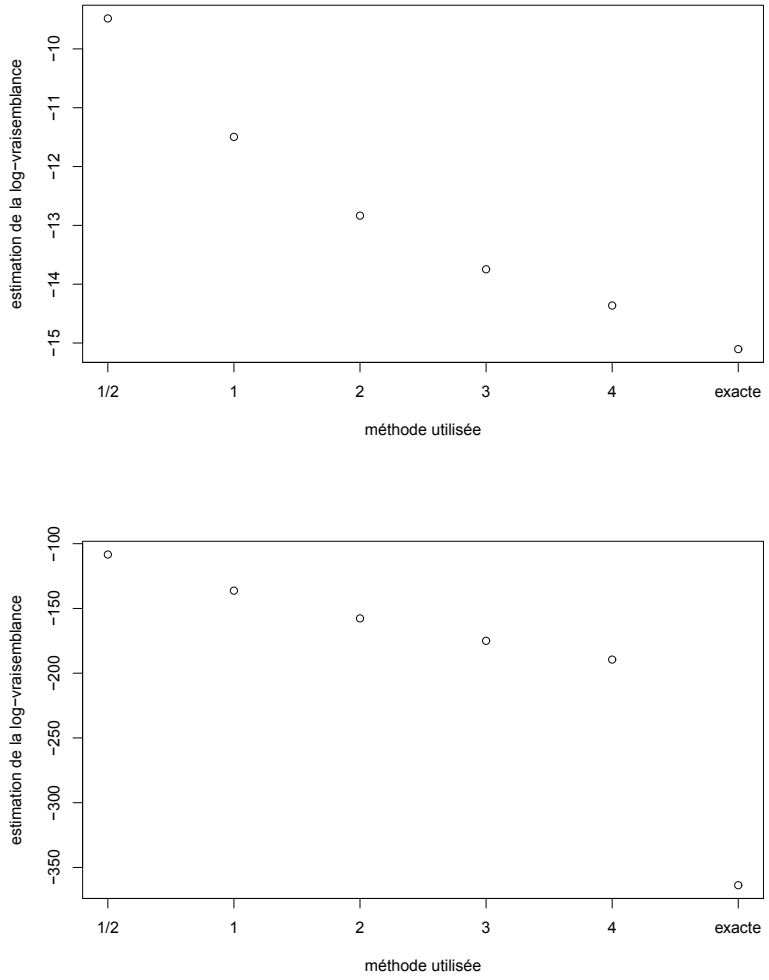


FIGURE 5.2 – Pour les séquences de longueurs 10 et 100 issues de l'exemple 5.1.10, log-vraisemblances estimées par $\hat{L}_{k\text{-Markov}}$ avec $k \in \{1/2, 1, 2, 3, 4\}$ et log-vraisemblance exacte.

$$\underbrace{G \dots G}_{m-1} A. \tag{5.6}$$

On calcule la probabilité d'avoir C à la racine au site 1, d'une part sachant que la séquence observée est (5.5), d'autre part sachant que la séquence observée est (5.6). On compte pour cela le nombre de permutations permettant d'obtenir la séquence observée souhaitée :

Proposition 5.1.12. *Le nombre de permutations correspondant aux évolutions :*
- *de la racine $CC \dots C$ à la séquence observée $G \dots GG$ est 1 (dans $\mathfrak{S}_m$).*
- *de la racine $CC \dots C$ à la séquence observée $G \dots GA$ est $m-1$ (dans $\mathfrak{S}_m$).*
- *de la racine $GC \dots C$ à la séquence observée $G \dots GG$ est 1 (dans $\mathfrak{S}_{m-1}$).*
- *de la racine $GC \dots C$ à la séquence observée $G \dots GA$ est $m-2$ (dans $\mathfrak{S}_{m-1}$).*

Démonstration. On utilise la propriété 5.1.5. Pour le premier point, une seule permutation de $\mathfrak{S}_m$ a $m-1$ montées. Pour le deuxième point, on a $\sigma(m-1)$ maximum global donc $\sigma(m-1) = m$. Il reste $m-1$ choix pour l'image de $\sigma(m)$ et cela détermine entièrement σ. On raisonne de même pour le troisième et le quatrième points. □

Conséquence 5.1.13. *Pour $m \geq 2$, la probabilité d'avoir C à la racine au site 1*
- *sachant que la séquence observée est $G \dots GG$ est $\frac{1}{m+1}$.*
- *sachant que la séquence observée est $G \dots GA$ est $\frac{m-1}{m^2-m-1}$.*

L'écart entre ces deux probabilités est équivalent à $\frac{1}{m^2}$.

Démonstration. Comme la loi à la racine est uniforme parmi deux séquences, la première probabilité s'exprime sous la forme :

$$\frac{1/m!}{1/m! + 1/(m-1)!}$$

et la deuxième sous la forme :

$$\frac{(m-1)/m!}{(m-1)/m! + (m-2)/(m-1)!}.$$

Le calcul direct de la différence conclut la démonstration de l'énoncé. □

Deuxième exemple.

On considère le même modèle, la même loi à la racine que pour le premier exemple (le modèle de l'exemple 5.1.1 avec comme loi à la racine la loi uniforme parmi les deux séquences (5.1) et (5.2)), et on considère la séquence observée suivante :

$$G \underbrace{GAGAGA \dots}_{m-1}. \tag{5.7}$$

On compte à nouveau le nombre de permutations permettant d'obtenir la séquence observée souhaitée. Ce nombre fait intervenir la notion de permutations alternantes.

Définition 5.1.14. *Une permutation σ est alternante si : $\sigma(1) > \sigma(2) < \sigma(3) > \dots$.*

On en déduit la proposition 5.1.15.

Proposition 5.1.15. *Le nombre de permutations correspondant aux évolutions :*
- *de la racine $CC\dots C$ à la séquence observée (5.7) est le nombre de permutations alternantes de $\mathfrak{S}_m$.*
- *de la racine $GC\dots C$ à la séquence observée (5.7) est le nombre de permutations alternantes de $\mathfrak{S}_{m-1}$.*

L'évolution du nombre de permutations alternantes E_m en fonction du nombre de lettres m à permuter a été étudié par André au XIXe siècle et est égal aux coefficients de la série génératrice exponentielle de $x \mapsto sec(x) + tan(x)$ [3, 108]. On obtient par conséquent la proposition suivante :

Proposition 5.1.16. *Pour $m \geq 2$, la probabilité d'avoir C à la racine au site 1 sachant que la séquence observée est (5.7) est :*

$$\frac{E_m/m!}{E_m/m! + E_{m-1}/(m-1)!}.$$

Cette suite est oscillante et converge vers $\frac{2}{2+\pi} \approx 0.39$ (voir la figure 5.3).

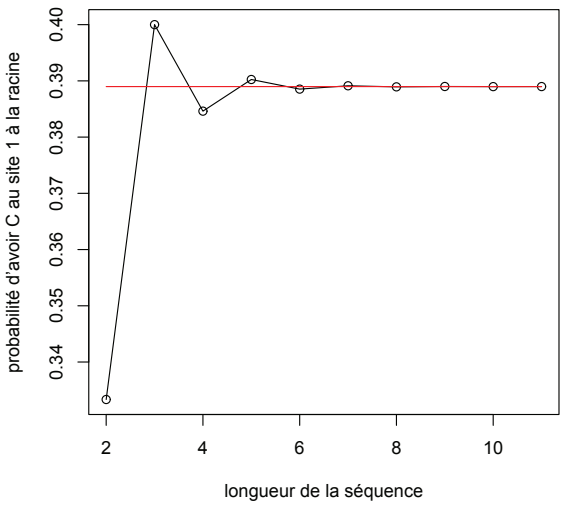


FIGURE 5.3 – Probabilité d'obtenir C à la racine du site 1 en fonction de la longueur de la séquence dans le deuxième exemple. La droite rouge correspond à la constante $\frac{2}{2+\pi}$.

5.2 Loi stationnaire de l'approximation markovienne

Dans cette section, on considère certains modèles d'évolution pour lesquels on compare les lois stationnaires obtenues par approximations markoviennes pour différents pas l. En particulier, on cherche à quantifier l'influence des nucléotides précédant le dinucléotide encodé considéré pour l'estimation de la loi stationnaire.

Pour cela, on introduit dans la section 5.2.1 un écart de probabilités $\Delta^l(\mathsf{M})$ (pour $l = 1/2$ ou $l \geq 1$) permettant de quantifier la différence d'estimation entre la loi stationnaire obtenue par l'approximation markovienne à l pas et l'approximation markovienne à $l+1$ pas.

On cherche par la suite à calculer ces quantités. Dans la section 5.2.2, la quantité $\Delta^{1/2}$ est calculée théoriquement pour deux modèles d'évolution extrêmes, pour lesquels on montre une forte influence des nucléotides précédant le dinucléotide considéré pour l'estimation de la loi stationnaire. Les quantités Δ^l (pour $l \in [\![1, 4]\!]$) sont ensuite calculés numériquement.

Dans la section 5.2.3, on considère des modèles M tirés aléatoirement et on calcule numériquement la quantité $\Delta^{1/2}(\mathsf{M})$. On montre que l'influence des nucléotides précédant le dinucléotide considéré pour l'estimation de la loi stationnaire est souvent faible mais présente.

On utilise la notation suivante :

Notation 5.2.1. *Pour un dinucléotide encodé $z_{i-1} = (r_{i-1}, n_i)$, on note $p_i = \pi(n_i)$.*

5.2.1 Définition des quantités recherchées

On considère l'approximation markovienne à l pas (pour $l = 1/2$ ou $l \geq 1$) et un modèle d'évolution M. On choisit un modèle d'évolution de séquence à séquence, pour une longueur temporelle $T \to +\infty$. Lorsque le modèle considéré est irréductible et apériodique, la probabilité stationnaire d'être égal au dinucléotide encodé z_i au site i sachant les dinucléotides précédents $z_1, \ldots, z_{i-1}$ est donnée par :

$$\begin{aligned}
&\chi_{1/2}(p_i, z_i) = P(Z_i(T) = z_i \mid \pi(X_i(T)) = p_i) && \text{pour } l = 1/2,\\
&\chi_l(z_{i-l} \ldots z_{i-1}, z_i) = P(Z_i(T) = z_i \mid Z_{i-1}(T) = z_{i-1}, \ldots, Z_{i-l}(T) = z_{i-l}) && \text{pour } l \geq 1.
\end{aligned}$$

Ainsi, on est amenés à calculer

$$\begin{aligned}
&P(\pi(X_i(T)) = p_i),\\
&P(Z_{i-l}(T) = z_{i-l}, \ldots, Z_{i-1}(T) = z_{i-1}, Z_i(T) = z_i) \text{ (pour } l \geq 1),
\end{aligned}$$

que l'on obtient à l'aide des exponentielles de matrices associées, de tailles resp. 2×2, $9.4^{l-1} \times 9.4^{l-1}$ (pour $l \geq 1$).

Pour chaque modèle M, on quantifie alors l'influence des nucléotides précédant le dinucléotide encodé considéré pour l'estimation de la loi stationnaire grâce aux valeurs suivantes :

$$\begin{aligned}
\Delta^{1/2}(\mathsf{M}) &= \max_{(z_{i-1}, z_i)} |\chi_1(z_{i-1}, z_i) - \chi_{1/2}(p_i, z_i)|,\\
\Delta^l(\mathsf{M}) &= \max_{(z_{i-l-1} \ldots, z_i)} |\chi_{l+1}(z_{i-l-1} \ldots z_{i-1}, z_i) - \chi_l(z_{i-l} \ldots z_{i-1}, z_i)| \qquad \text{pour } l \geq 1.
\end{aligned}$$

Notons que dans les cas où le modèle d'évolution n'est pas irréductible mais possède une seule classe récurrente, on considère la loi stationnaire sur cette classe restreinte.

5.2.2 Un modèle extrême

On considère un modèle extrême où les quantités peuvent être calculées théoriquement. Ce modèle est construit à partir du modèle extrême limite de la section 5.1, où on a rajouté différents taux pour rendre l'évolution davantage symétrique.

Exemple 5.2.2. *On définit les modèles d'évolution* $\mathsf{M}_{extIrr}(\varepsilon)$ *(pour* $\varepsilon > 0$*), où les coefficients non indiqués sont nuls :*

$$v_C = v_G = 1, \quad r_{TG\to CG} = r_{CG\to CA} = r_{CA\to TA} = 1/\varepsilon.$$

$$v_A = v_T = w_A = w_C = w_G = w_T = \varepsilon.$$

Le modèle limite quand $\varepsilon \to 0$ *est noté* $\mathsf{M}_{extIrr}(0)$ *et est défini par :*

$$v_C = v_G = 1, \quad r_{TG\to CG} = r_{CG\to CA} = r_{CA\to TA} \to +\infty.$$

On représente sur la figure 5.5 le graphe associé à la matrice d'évolution des couples encodés du modèle $\mathsf{M}_{extIrr}(0)$ (les dinucléotides se substituant instantanément ne sont pas indiqués). On remarque que ce modèle limite est irréductible pour les couples encodés mais pas pour les triplets encodés (en effet, on ne passe jamais par TGG). Par comparaison, on représente sur la figure 5.4 le graphe associé à la matrice d'évolution des couples encodés du modèle extrême limite décrit dans la section 5.1 (non irréductible).

On étudie maintenant la loi stationnaire du modèle limite $\mathsf{M}_{extIrr}(0)$.

Calcul théorique de $\Delta^{1/2}$ pour le modèle $\mathsf{M}_{extIrr}(0)$.

Calculs de $\chi_{1/2}$. Pour les nucléotides encodés dans π, on sait que $P(R) = \frac{v_A+v_G}{v_A+v_C+v_G+v_T}$ et $P(Y) = 1 - P(R)$. Ainsi :

$$P(R) = 1/2 \text{ et } P(Y) = 1/2.$$

Pour les dinucléotides encodés, la matrice d'évolution des couples encodés s'écrit :

$$\begin{array}{c} \\ RY \\ TA \\ RA \\ TY \\ RG \\ CY \end{array} \begin{array}{c} \begin{array}{cccccc} RY & TA & RA & TY & RG & CY \end{array} \\ \begin{pmatrix} . & 0 & 0 & 0 & 1 & 1 \\ 0 & . & 1 & 1 & 0 & 0 \\ 1 & 1 & . & 0 & 0 & 0 \\ 1 & 1 & 0 & . & 0 & 0 \\ 1 & 1 & 0 & 0 & . & 0 \\ 1 & 1 & 0 & 0 & 0 & . \end{pmatrix} \end{array}.$$

En regroupant les dinucléotides qui ont le même comportement, d'une part RY et TA, d'autre part RA, TY, RG et CY, on obtient la matrice

$$\begin{array}{c} \\ \{RY, TA\} \\ \{RA, TY, RG, CY\} \end{array} \begin{array}{c} \begin{array}{cc} \{RY, TA\} & \{RA, TY, RG, CY\} \end{array} \\ \begin{pmatrix} . & 2 \\ 2 & . \end{pmatrix} \end{array}.$$

On a donc $P(RA) = \frac{P(\{RA,TY,RG,CY\})}{4} = 1/8$ et :

$$\begin{aligned} P(RY) = P(TA) &= 1/4, \\ P(RA) = P(RG) = P(CY) = P(TY) &= 1/8, \\ P(CA) = P(CG) = P(TG) &= 0. \end{aligned}$$

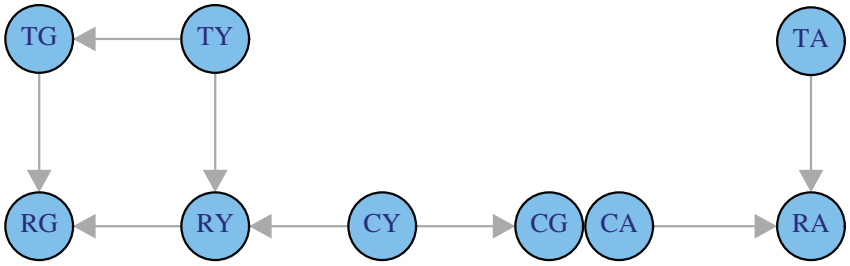


FIGURE 5.4 – Graphe associé à la matrice d'évolution des couples encodés pour le modèle limite décrit dans la section 5.1. Chaque arête orientée a un taux de 1, sauf l'arête $CA \to CG$ qui a un taux infini.

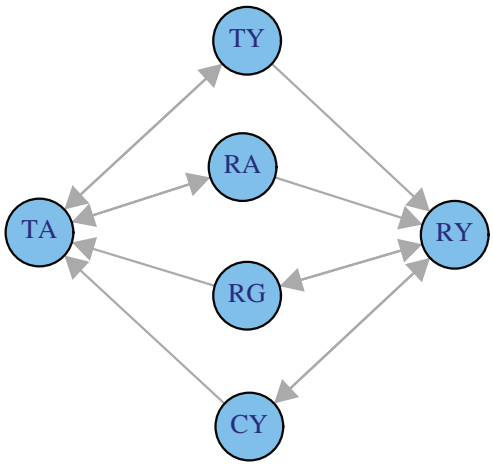


FIGURE 5.5 – Graphe associé à la matrice d'évolution des couples encodés pour le modèle $\mathsf{M}_{extIrr}(0)$.

On en déduit les valeurs de $\chi_{1/2}$:

$$\begin{aligned}\chi_{1/2}(R,RA)=1/4,\ \chi_{1/2}(R,RG)=1/4,\ \chi_{1/2}(R,RY)=1/2,\\ \chi_{1/2}(Y,CA)=0,\ \chi_{1/2}(Y,CG)=0,\ \chi_{1/2}(Y,CY)=1/4,\\ \chi_{1/2}(Y,TA)=1/2,\ \chi_{1/2}(Y,TG)=0,\ \chi_{1/2}(Y,TY)=1/4.\end{aligned}$$

Calculs de χ_1. Pour obtenir $\Delta^{1/2}$, il reste à calculer la probabilité des différents triplets encodés. On procède comme pour les dinucléotides encodés et on obtient la proposition suivante :

Proposition 5.2.3. *On fait les regroupements suivants :*

$$\begin{aligned}\{RTA, TAY\}&\ \textit{représenté par } RTA,\\ \{RGA, TCY\}&\ \textit{représenté par } RGA,\\ \{RGY, RCY, TAA, TTA\}&\ \textit{représenté par } RGY,\\ \{CTY, RAG\}&\ \textit{représenté par } CTY,\\ \{RTY, CTA, RAY, TAG\}&\ \textit{représenté par } RTY,\\ \{RAA, RGG, CCY, TTY\}&\ \textit{représenté par } RAA.\end{aligned}$$

La matrice d'évolution dans ces classes s'écrit alors :

$$M = \begin{array}{c} \\ RTA \\ RGA \\ RGY \\ CTY \\ RTY \\ RAA \end{array} \begin{array}{c} \begin{array}{cccccc} RTA & RGA & RGY & CTY & RTY & RAA \end{array} \\ \begin{pmatrix} . & 1 & 0 & 0 & 2 & 0 \\ 1 & . & 2 & 0 & 0 & 0 \\ 1 & 0 & . & 0 & 0 & 1 \\ 1 & 0 & 0 & . & 2 & 0 \\ 1 & 0 & 1 & 1 & . & 0 \\ 1 & 0 & 1 & 0 & 1 & . \end{pmatrix} \end{array}.$$

On en déduit que la probabilité stationnaire d'être dans ces classes est respectivement : 1/4, 1/12, 1/4, 1/12, 1/4 *et* 1/12.

Démonstration. La matrice d'évolution est donnée par :

	RTA	TAY	RGA	TCY	RGY	RCY	TAA	TTA	CTY	RAG	RTY	CTA	RAY	TAG	RAA	RGG	CCY	TTY
RTA	.	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
TAY	0	.	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0
RGA	1	0	.	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
TCY	0	1	0	.	0	1	0	1	0	0	0	0	0	0	0	0	0	0
RGY	0	1	0	0	.	1	0	0	0	0	0	0	0	0	0	1	0	0
RCY	1	0	0	0	1	.	0	0	0	0	0	0	0	0	0	0	1	0
TAA	0	1	0	0	0	0	.	1	0	0	0	0	0	0	1	0	0	0
TTA	1	0	0	0	0	0	1	.	0	0	0	0	0	0	0	0	0	1
CTY	0	1	0	0	0	0	0	0	.	0	1	1	0	0	0	0	0	0
RAG	1	0	0	0	0	0	0	0	0	.	0	0	1	1	0	0	0	0
RTY	1	0	0	0	1	0	0	0	1	0	.	0	0	0	0	0	0	0
CTA	1	0	0	0	0	0	1	0	1	0	0	.	0	0	0	0	0	0
RAY	0	1	0	0	0	1	0	0	0	1	0	0	.	0	0	0	0	0
TAG	0	1	0	0	0	0	0	1	0	1	0	0	0	.	0	0	0	0
RAA	1	0	0	0	0	0	1	0	0	0	0	0	1	0	.	0	0	0
RGG	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	.	0	0
CCY	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	.	0
TTY	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	.

.

On souhaite regrouper RTA et TAY qui ont un comportement similaire en $\{RTA, TAY\}$ (1). Cela impose le regroupement $\{RGA, TCY\}$ (2) et $\{RTY, CTA, RAY, TAG\}$ (5). Le regroupement (2) impose les regroupements d'une part $\{RGY, RCY\}$ (3a) et d'autre part $\{TAA, TTA\}$ (3b). Le regroupement (5) impose le regroupement $\{CTY, RAG\}$ (4) et de regrouper ensemble (3a) et (3b), noté alors (3). Le regroupement (3) impose le regroupement $\{RAA, RGG, CCY, TTY\}$ (6).

On vérifie que ces regroupements sont possibles et on obtient alors la matrice M voulue.

La matrice obtenue se décompose sous forme de Jordan $M = PJP^{-1}$, avec :

$$J = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -3 & 0 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 & 0 \\ 0 & 0 & 0 & -4 & 1 & 0 \\ 0 & 0 & 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & 0 & 0 & -4 \end{pmatrix},$$

$$P = \frac{1}{48}\begin{pmatrix} 1 & 32 & 0 & -12 & -33 & 0 \\ 1 & 32 & -6 & 0 & 21 & 24 \\ 1 & -16 & -3 & 6 & 6 & -12 \\ 1 & -16 & 6 & 0 & 33 & 24 \\ 1 & -16 & 3 & 6 & 0 & -12 \\ 1 & -16 & 0 & 0 & 27 & 24 \end{pmatrix} \text{ et } P^{-1} = \begin{pmatrix} 12 & 4 & 12 & 4 & 12 & 4 \\ 0 & 1 & 0 & 1 & 0 & -2 \\ 0 & 0 & -4 & 4 & 4 & -4 \\ -3 & 3 & -10 & -8 & 12 & 6 \\ 0 & 0 & 4 & 4 & -4 & -4 \\ -1/2 & 1/2 & -5 & -4 & 4 & 5 \end{pmatrix}.$$

On a ensuite

$$\exp(TJ) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

d'où on déduit la loi stationnaire voulue. □

De la proposition 5.2.3 on déduit :

$$\begin{aligned} P(RTA) = P(TAY) &= 1/8, \\ P(RGA) = P(TCY) = P(CTY) = P(RAG) &= 1/24, \\ P(RGY) = P(RCY) = P(TAA) = P(TTA) &= 1/16, \\ P(RTY) = P(CTA) = P(RAY) = P(TAG) &= 1/16, \\ P(RAA) = P(RGG) = P(CCY) = P(TTY) &= 1/48, \end{aligned}$$

puis les valeurs de χ_1, en particulier :

$$\begin{aligned} &\chi_1(RA,RA) = 1/6,\ \chi_1(RG,RA) = 1/3,\ \chi_1(TA,RA) = 1/4, \\ &\chi_1(RA,RG) = 1/3,\ \chi_1(RG,RG) = 1/6,\ \chi_1(TA,RG) = 1/4, \\ &\chi_1(RY,CY) = 1/4,\ \chi_1(CY,CY) = 1/6,\ \chi_1(TY,CY) = 1/3, \\ &\chi_1(RY,TY) = 1/4,\ \chi_1(CY,TY) = 1/3,\ \chi_1(TY,TY) = 1/6. \end{aligned}$$

Obtention de $\Delta^{1/2}(\mathsf{M}_{extIrr}(0))$. On obtient enfin, d'après les calculs de $\chi_{1/2}$ et χ_1 :

$$\Delta^{1/2}(\mathsf{M}_{extIrr}(0)) = 1/12 \approx 8\%.$$

Calcul théorique de $\Delta^{1/2}$ pour le modèle $\mathsf{M}_{extIrr}(\varepsilon)$ quand $\varepsilon \to 0$.

On va maintenant comparer la valeur $\Delta^{1/2}(\mathsf{M}_{extIrr}(0))$ avec celle de $\Delta^{1/2}(\mathsf{M}_{extIrr}(\varepsilon))$ pour $\varepsilon \to 0$. Les deux valeurs sont a priori différentes car le calcul de $\Delta^{1/2}(\mathsf{M}_{extIrr}(\varepsilon))$ fait intervenir les triplets encodés débutant par CA, CG ou TG, ce qui n'est plus le cas dans le modèle $\mathsf{M}_{extIrr}(0)$.

Proposition 5.2.4. *On a :* $\Delta^{1/2}(\mathsf{M}_{extIrr}(\varepsilon)) \to 1/4$.

Démonstration. On sait par le paragraphe précédent que $\lim \Delta^{1/2}(\mathsf{M}_{extIrr}(\varepsilon)) \geq 1/12$, et que :

$$\chi_{1/2}(R, RA) = 1/4, \quad \chi_{1/2}(R, RG) = 1/4, \quad \chi_{1/2}(R, RY) = 1/2.$$

Par le corollaire 3.5.2 de découpage RY, on a

$$\chi_1(z_{i-1}, RY) = 1/2$$

pour $z_{i-1} \in \{CA, CG, TG\}$. Cela implique que $\chi_1(z_{i-1}, z_i) \leq 1/2$ pour tout triplet vérifiant $z_{i-1} \in \{CA, CG, TG\}$, et donc que :

$$\Delta^{1/2}(\mathsf{M}_{extIrr}(\varepsilon)) \leq 1/4.$$

On calcule maintenant $\chi_1(TG, RA), \chi_1(TG, RG)$. On regarde depuis quels triplets on peut accéder à TGA et TGG :

- TGA ne peut provenir (avec probabilité non négligeable) que de TTA ou de TCA. Mais la probabilité d'être issu de TCA est aussi négligeable puisque $r_{CA\to TA} = 1/\varepsilon$. Ainsi TGA est issu de TTA avec une probabilité non négligeable.
- TGG ne peut provenir (avec probabilité non négligeable) que de TTG, TCG ou TGY. Mais les probabilités d'être issu de ces triplets est aussi négligeable.

Ainsi, $\chi_1(TG, RG) \to 0$, $\chi_1(TG, RA) \to 1/2$, puis $|\chi_1(TG, RG) - \chi_{1/2}(R, RG)| \to 1/4$. □

Calculs numériques de $\Delta^l(\mathsf{M}_{extIrr}(\varepsilon))$ pour $\varepsilon \to 0$ et $l \geq 1$.

Les résultats sont obtenus numériquement et regroupés dans la table 5.1.

Δ^l	% obtenu	séquence où est atteint le max
1/2	25	TGA
1	12.5	CCAG
2	7.25	CCGGA
3	3.42	CCGGAA
4	2.07	CCGGAAG

TABLE 5.1 – Calculs de $\Delta^l(\mathsf{M}_{extIrr}(\varepsilon))$ pour $\varepsilon \to 0$ et $l \in [\![1, 4]\!]$.

5.2.3 Modèles simulés

On munit l'ensemble des paramètres du modèle RN95+YpR de la loi uniforme sur $[0,1]^{16}$. Pour un échantillon de modèles de taille un million, on calcule numériquement la quantité associée $\Delta_k^{1/2}$. Le modèle traité M_{ex1} est celui qui a obtenu la valeur la plus élevée :

$$\begin{aligned}
&v_A = 0.067, v_C = 0.0067, v_G = 0.10, v_T = 0.016,\\
&w_A = 0.40, w_C = 0.89, w_G = 0.0013, w_T = 0.82,\\
&r_{CG\to CA} = 0.60, r_{CA\to CG} = 0.92, r_{TA\to TG} = 0.77, r_{TG\to TA} = 0.02,\\
&r_{CA\to TA} = 0.45, r_{CG\to TG} = 0.16, r_{TA\to CA} = 0.12, r_{TG\to CG} = 0.21.
\end{aligned}$$

Les différentes valeurs de $\Delta^l(\mathsf{M}_{ex1})$ pour $l \in [\![1,4]\!]$ sont regroupés dans la table 5.2.

Δ^l	% obtenu	séquence où est atteint le max
1/2	9.5	RGA
1	7.0	RGGA
2	3.5	RGGGA
3	2.3	RGGGGA
4	1.3	RGGGGGA

TABLE 5.2 – Calculs de $\Delta^l(\mathsf{M}_{ex1})$ pour $l \in [\![1,4]\!]$.

Comportement moyen. On reprend l'échantillon de l'exemple précédent et on regarde la fonction de répartition de $(\log \Delta_k^{1/2})_{k\in[\![1,1000000]\!]}$. On observe sur la figure 5.6 que 95% des modèles considérés ont un écart compris entre $\exp(-10.36) = 3.10^{-5}$ et 9.10^{-3}. L'écart est donc souvent faible mais strictement positif.

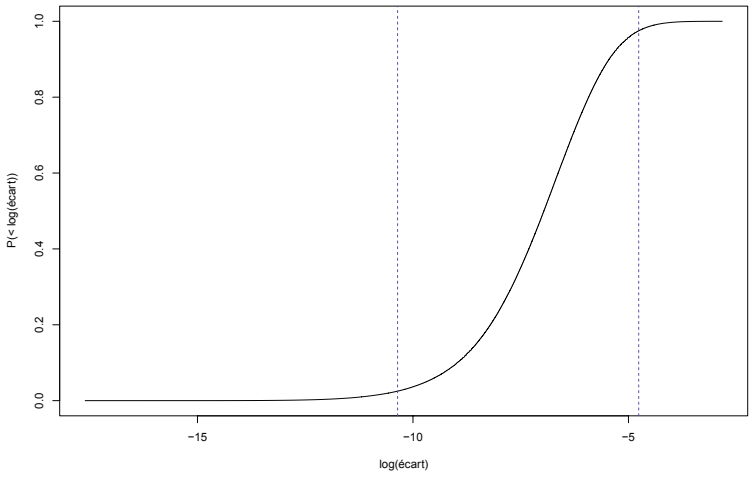


FIGURE 5.6 – Fonction de répartition empirique de la quantité $\log \Delta^{1/2}$ pour des modèles RN95+YpR tirés uniformément sur $[0,1]^{16}$.

Chapitre 6

Structures markoviennes

Ce chapitre est consacré à l'étude de la dépendance spatiale existant entre les histoires évolutives aux différents sites des modèles RN95+YpR. Les structures markoviennes qui en résultent constituent la base des méthodes de simulation développées aux chapitres suivants.

On considère un processus de Markov (par rapport à la variable t)

$$X = (X(t))_{t\in[0,T]} = (X_1(t), \ldots, X_m(t))_{t\in[0,T]}$$

décrivant l'évolution temporelle d'une séquence de m sites consécutifs, et on cherche à comprendre la dépendance entre les $X_i := (X_i(t))_{t\in[0,T]}$ pour les différentes valeurs de i.

Pour un modèle markovien avec dépendance au voisin immédiat général, $(X_i)_{i\in[\![1,m]\!]}$ possède une structure de champ markovien d'ordre 2, avec une description explicite (section 6.1.1). Lorsque le modèle markovien appartient à la classe RN95+YpR, on montre que cette structure devient une structure de champ markovien d'ordre 1, avec une description également explicite (sections 6.1.2, 6.1.3 et 6.1.4). Chacune de ces structures de champ markovien induit une structure associée de chaîne de Markov d'ordre un par rapport aux sites, mais dont la description n'est plus explicite à cause d'intégrations successives à effectuer (section 6.1.5). Par conséquent, il est possible de faire de l'échantillonnage de Gibbs à partir de ces structures de champ markovien, mais pas de simuler la structure correspondante de chaîne de Markov à la base des méthodes particulaires.

On cherche alors à obtenir une structure spatiale markovienne de l'évolution dont les transitions s'expriment explicitement (section 6.2). Pour cela, on définit pour chaque entier $i \in [\![1, m-1]\!]$ les évolutions des dinucléotides encodés avec chevauchement :

$$Z_i := (\rho(X_i), \eta(X_{i+1}))$$

puis on identifie la séquence $(\rho(X_1), X_2, \ldots, X_{m-1}, \eta(X_m))$ avec $(Z_1, Z_2, \ldots, Z_{m-1}) = (Z_i)_{i\in[\![1,m-1]\!]}$. On établit que la suite $(Z_i)_{i\in[\![1,m-1]\!]}$ est une chaîne de Markov à un pas dont on explicite le comportement. La simulation de cette chaîne de Markov rend par la suite possible l'utilisation des méthodes particulaires.

Dans la section 6.3, on décrit une structure d'indépendance conditionnelle à l'évolution π-encodée, qui n'est pas à proprement parler une structure markovienne mais qui est placée ici par commodité.

Enfin, on généralise pour un arbre phylogénétique quelconque et une loi à la racine générale ces structures initialement énoncées pour la topologie de séquence à séquence et une racine fixée (section 6.4).

Notations pour ce chapitre.

Notation 6.0.5. *On utilise les notations suivantes pour les matrices de taux de sauts instantanées décrivant l'évolution du premier site (définition 3.2.5), des sites intermédiaires (définition 1.2.5 et remarque 1.2.6) et du dernier site (définition 3.2.6) d'une séquence Φ-encodée. Pour $g, d \in \mathcal{A}$:*

$$Q_{\rho(g),\eta(d)} := Q_{g,d},\ Q_{,d} := Q_{,\eta(d)} \text{ et } Q_{g,} := Q_{\rho(g),}.$$

Aussi, pour $N, N' \in \mathcal{A}$,

$$Q_{,d}(N, N') := Q_{,d}(\rho(N), \rho(N')) \text{ et } Q_{g,}(N, N') := Q_{g,}(\eta(N), \eta(N')).$$

Enfin, pour tout $\varepsilon > 0$ on écrit $q^{\varepsilon}_{g,d}, q^{\varepsilon}_{,d}$ et $q^{\varepsilon}_{g,}$ les matrices de transition instantanée (voir définition 1.5.6) associées aux matrices de taux de sauts $Q_{g,d}, Q_{,d}$ et $Q_{g,}$. Un élément de cette matrice est appelé taux de substitution instantané.

6.1 Structure spatiale de champ markovien

On va établir dans cette section que les évolutions issues de la classe RN95+YpR vérifient une propriété spatiale de champ markovien d'ordre un de chaque site par rapport à ses voisins immédiats, explicite en les paramètres du modèle. On verra aussi que l'on obtient la même structure en conditionnant l'évolution par la donnée d'une séquence associée à l'heure actuelle. On considère dans cette section uniquement une évolution de séquence à séquence. Les mêmes propriétés seront adaptées sur un arbre dans la section 6.4.

Les notations suivantes sont utilisées pour simplifier l'écriture des calculs.

Notation 6.1.1. *Pour tout site j et tous instants $s < t$, on écrit :*

- *$x_j(s \to t)$ une évolution au site j de s à t (l'espace d'état associé est l'ensemble des évolutions càdlàg avec un nombre fini de sauts entre les instants s et t et de longueur* 1, *la description étant analogue à celle donnée dans la section 1.5).*
  On utilise ensuite :
- *$x_j(s)$ à la place de $X_j(s) = x_j(s)$ dans l'écriture des probabilités.*
- *$x_j(s \to t)$ à la place de $(X_j(u))_{u \in [s,t]} = x_j(s \to t)$.*

6.1.1 Structure de champ markovien d'ordre deux

Dans cette section, on formalise un modèle d'évolution plus général que RN95+YpR dans lequel les taux de substitution en un site peuvent dépendre de manière arbitraire des voisins immédiats. Pour ce modèle d'évolution avec dépendance aux voisins immédiats

quelconque, on obtient que la séquence évolutive est un champ markovien d'ordre deux en les sites (adapté ici de [27, 28]).

Pour cela, on écrit d'abord dans la définition 6.1.2 la forme de la matrice de taux de sauts de séquence à séquence dans le cas d'un modèle d'évolution avec dépendance aux voisins immédiats. Ensuite, on utilise l'expression de la densité des évolutions (énoncée dans proposition 1.5.4) pour montrer la structure de champ markovien de l'évolution (théorème 6.1.3).

On note $x = x_{1:m}$ une évolution sur $[0,T]$ d'une séquence de longueur m, et on pose de façon arbitraire $x_{-1} = x_0 = x_{m+1} = x_{m+2} \equiv A$. Une séquence de $\mathcal{A}^m$ est notée de façon générique $x(t) = (x_1(t), \ldots, x_m(t))$. Le marqueur temporel t sert ici à ne pas confondre une séquence avec une évolution $x \in \mathsf{E}$ (où l'ensemble des évolutions E est défini dans la section 1.5).

Définition 6.1.2. *Pour $(g,d) \in \mathcal{A}^2$, soit $\boldsymbol{Q}_{g,d}$ une matrice de taux de sauts sur $\mathcal{A}$.*

On dit que l'évolution de séquence à séquence sur l'intervalle $[0,T]$ régie par la matrice de taux de sauts $Q_{séq}$ est à dépendance aux voisins immédiats associée aux matrices de taux de sauts $\boldsymbol{Q}_{g,d}$ si,

- *pour toutes séquences de $\mathcal{A}^m$ ne différant que d'un nucléotide*

  $$x(t) = (x_1(t), \ldots, x_m(t)) \quad \text{et} \quad \tilde{x}(t) = (x_1(t), \ldots, x_{i-1}(t), \tilde{x}_i(t), x_{i+1}(t), \ldots, x_m(t)),$$

  la matrice de taux de sauts $Q_{séq}$ vérifie :

  $$Q_{séq}(x(t), \tilde{x}(t)) = \boldsymbol{Q}_{x_{i-1}(t), x_{i+1}(t)}(x_i(t), \tilde{x}_i(t)).$$

- *pour toutes séquences de $\mathcal{A}^m$ différant de plus d'un nucléotide, le taux de saut associé est nul.*

Tous les sauts de séquence à séquence d'une évolution à dépendance aux voisins immédiats sont donc déterminés par les différentes matrices $\boldsymbol{Q}_{g,d}$ ($(g,d) \in \mathcal{A}^2$). On établit alors une propriété de champ markovien d'ordre deux d'une telle évolution :

Théorème 6.1.3. *On suppose que l'évolution sur l'intervalle $[0,T]$ est régie par la matrice de taux de sauts $Q_{séq}$, à dépendance aux voisins immédiats associée aux matrices de taux de sauts $\boldsymbol{Q}_{g,d}$ (avec $(g,d) \in \mathcal{A}^2$).*

On choisit comme état initial $x(0) = (x_1(0), \ldots, x_m(0)) \in \mathcal{A}^m$ et la chaîne de Markov (dans le temps) associée à cette évolution

$$X = (X(t))_{t\in[0,T]} = (X_1(t), \ldots, X_m(t))_{t\in[0,T]}\,.$$

On note en considérant l'évolution site par site $X = (X_i)_{i\in[\![1,m]\!]}$.

Alors $(X_i)_{i\in[\![1,m]\!]}$ est un champ markovien d'ordre 2 (c'est-à-dire que l'évolution du site i conditionnellement aux évolutions $X_{1:i-1}$ et $X_{i+1:m}$ ne dépend que de X_{i-2}, X_{i-1}, X_{i+1} et X_{i+2}).

Cas discret. Pour donner une idée de la preuve, on montre d'abord le résultat lorsque la variable temporelle est discrète. L'intervalle $[0,T]$ est alors remplacé par l'ensemble $[\![0,T]\!]$.

On suppose connaître un noyau de transition $q_{\text{séq}}$ sur $\mathcal{A}^m$ et des noyaux de transitions $q_{g,d}$ sur $\mathcal{A}$ (pour $(g,d) \in \mathcal{A}^2$) vérifiant pour toutes séquences $x(t-1)$ et $x(t)$ de $\mathcal{A}^m$:

$$q_{\text{séq}}(x(t-1), x(t)) = \prod_{k=1}^{m} q_{x_{k-1}(t-1), x_{k+1}(t-1)}(x_k(t-1), x_k(t)).$$

On fixe maintenant un site $i \in [\![1, m]\!]$. Pour une évolution $x = (x_k)_{k \in [\![1,m]\!]}$ de probabilité strictement positive, on exprime la probabilité conditionnelle $P(x_i | x_{1:i-1}, x_{i+1:m})$ de la façon suivante :

$$P := P(x_i | x_{1:i-1}, x_{i+1:m}) = \frac{P(x)}{\sum_{\tilde{x}_i} P(\tilde{x})},$$

en ayant noté pour chaque évolution $\tilde{x}_i$ du site i :

$$\tilde{x} := (\tilde{x}_1, \ldots, \tilde{x}_m) := (x_1, \ldots, x_{i-1}, \tilde{x}_i, x_{i+1}, \ldots, x_m).$$

Or, le numérateur s'exprime comme :

$$P(x) = \prod_{t=1}^{T} q_{\text{séq}}(x(t-1), x(t)) = \prod_{t=1}^{T} \prod_{k=1}^{m} q_{x_{k-1}(t-1), x_{k+1}(t-1)}(x_k(t-1), x_k(t)).$$

En écrivant de la même manière le dénominateur $P(\tilde{x})$ pour chaque $\tilde{x}_i$, le quotient P s'exprime alors, en simplifiant les termes qui ne dépendent pas de $\tilde{x}_i$, par :

$$P = \frac{\prod_{t=1}^{T} \prod_{k=i-1}^{i+1} q_{x_{k-1}(t-1), x_{k+1}(t-1)}(x_k(t-1), x_k(t))}{\sum_{\tilde{x}_i} \prod_{t=1}^{T} \prod_{k=i-1}^{i+1} q_{\tilde{x}_{k-1}(t-1), \tilde{x}_{k+1}(t-1)}(\tilde{x}_k(t-1), \tilde{x}_k(t))}$$

On remarque que P ne dépend de x qu'à travers $x_{i-2}, x_{i-1}, x_i, x_{i+1}$ et x_{i+2}, ce qui permet de conclure que l'évolution (discrète) est un champ markovien à deux pas.

Cas continu. Pour montrer le théorème 6.1.3, on cherche à adapter la preuve précédente lorsque l'on considère l'intervalle continu $[0, T]$. Par la proposition 1.5.4 et la notation 1.5.2, on sait exprimer pour chaque évolution

$$x = (l, (0 = t_0 < t_1 < \ldots < t_l < T), (x(t_0), \ldots, x(t_l)))$$

sa densité par rapport à la mesure de référence μ :

$$f(x) = \left[\prod_{k=0}^{l-1} e^{-(t_{k+1} - t_k) Q_{\text{séq}}(x(t_k))} Q_{\text{séq}}(x(t_k), x(t_{k+1})) \right] e^{-(T - t_l) Q_{\text{séq}}(x(t_l))}.$$

Il suffit alors de montrer que pour toute évolution $x = (x_1, \ldots, x_m)$ et pour tout site $i \in [\![1, m]\!]$, la densité conditionnelle :

$$f(x_i | x_{1:i-1}, x_{i+1:m}) := \frac{f(x)}{\int_{\tilde{x}_i} f(x_1, \ldots, x_{i-1}, \tilde{x}_i, x_{i+1}, \ldots, x_m) d\mu}$$

ne dépend pas du choix de $x_{1:i-3}$ et $x_{i+3:m}$.

Pour cela, on établit d'abord le lemme suivant :

Lemme 6.1.4. *Pour toute séquence $x(t) = x_{1:m}(t) \in \mathcal{A}^m$, pour tout site $i \in [\![1, m]\!]$, la quantité $Q_{seq}(x(t))$ (qui est définie par $-Q_{séq}(x(t), x(t)) > 0$ d'après la section 1.5) peut être décomposée en la somme de deux termes :*

- *$C_i(x_{1:i-1}(t), x_{i+1:m}(t))$ ne dépendant pas du nucléotide $x_i(t)$ considéré et*
- *$D_i(x_{i-2:i+2}(t))$ ne dépendant pas des nucléotides considérés en dehors des sites $i-2$ à $i+2$.*

En particulier, pour deux évolutions $x = (x_1, \ldots, x_m)$ et $\tilde{x} = (x_1, \ldots x_{i-1}, \tilde{x}_i, x_{i+1}, \ldots, x_m)$ ne différant qu'au site i et pour tout instant $t \in [0, T]$, la différence $Q_{séq}(x(t)) - Q_{séq}(\tilde{x}(t))$ ne dépend que des nucléotides $x_{i-2}(t), x_{i-1}(t), x_i(t), \tilde{x}_i(t), x_{i+1}(t)$ et $x_{i+2}(t)$.

Démonstration. Soit un site i et $x(t) = x_{1:m}(t) \in \mathcal{A}^m$. D'après la définition 6.1.2, les sauts accessibles à partir de la séquence $x(t)$ ne diffèrent de $x(t)$ qu'en un site. En parcourant chaque site et chaque saut possible, on peut donc écrire :

$$Q_{\text{seq}}(x(t)) = \sum_{j=1}^{m} \sum_{N \neq x_j(t)} \boldsymbol{Q}_{x_{j-1}(t), x_{j+1}(t)}(x_j(t), N).$$

On pose :

$$C_i(x_{1:i-1}(t), x_{i+1:m}(t)) = \sum_{j \neq \{i-1, i, i+1\}} \sum_{N \neq x_j(t)} \boldsymbol{Q}_{x_{j-1}(t), x_{j+1}(t)}(x_j(t), N)$$

et

$$D_i(x_{i-2:i+2}(t)) = \sum_{j=i-1}^{i+1} \sum_{N \neq x_j(t)} \boldsymbol{Q}_{x_{j-1}(t), x_{j+1}(t)}(x_j(t), N)$$

et on obtient $Q_{\text{seq}}(x(t)) = C_i(x_{1:i-1}(t), x_{i+1:m}(t)) + D_i(x_{i-2:i+2}(t))$ avec les conditions souhaitées. □

Passons maintenant à la démonstration du théorème 6.1.3.

Démonstration. On veut montrer que pour toute évolution $x = (x_1, \ldots, x_m)$ et pour tout site $i \in [\![1, m]\!]$, la densité conditionnelle :

$$f(x_i | x_{1:i-1}, x_{i+1:m}) := \frac{f(x)}{\int_{\tilde{x}_i} f(x_1, \ldots, x_{i-1}, \tilde{x}_i, x_{i+1}, \ldots, x_m) d\mu}$$

ne dépend pas du choix de $x_{1:i-3}$ et $x_{i+3:m}$.

On choisit alors $x = (x_1, \ldots, x_m)$ et $\tilde{x} = (x_1, \ldots x_{i-1}, \tilde{x}_i, x_{i+1}, \ldots, x_m)$ ne différant qu'au site i, et on cherche à exprimer le rapport de densités $\frac{f(\tilde{x})}{f(x)}$.

Comme pour toute évolution x et pour tout $t \in]0, T[$, l'égalité $f(x(0 \to T)) = f(x(0 \to t)) f(x(t \to T))$ est vérifiée (puisque l'évolution est markovienne en temps), il suit que pour tous les instants $s_0 := 0 < s_1 < \ldots < s_L < T$, le rapport de densités $\frac{f(\tilde{x})}{f(x)}$ s'exprime comme le produit des rapports de densités suivants (pour $k \in [\![0, L-1]\!]$) :

$$\frac{f(\tilde{x}(s_k \to s_{k+1}))}{f(x(s_k \to s_{k+1}))}.$$

On choisit les instants $s_1 < \ldots < s_L$ de telle sorte que sur chaque intervalle $[s_k, s_{k+1}]$, une seule substitution a eu lieu dans $x_1, \ldots x_m, \tilde{x}_i$ (dans le cas où aucune substitution n'a lieu globalement, alors $x = \tilde{x}$ et le rapport des densités est 1).

On choisit un intervalle $[s_k, s_{k+1}]$, on note s l'instant de la substitution. Le quotient de densité s'exprime alors d'après la proposition 1.5.4 par :

$$\frac{e^{-(s-s_k)Q_{\text{séq}}(\tilde{x}(s_k))}Q_{\text{séq}}(\tilde{x}(s_k),\tilde{x}(s_{k+1}))e^{-(s_{k+1}-s)Q_{\text{séq}}(\tilde{x}(s_{k+1}))}}{e^{-(s-s_k)Q_{\text{séq}}(x(s_k))}Q_{\text{séq}}(x(s_k),x(s_{k+1}))e^{-(s_{k+1}-s)Q_{\text{séq}}(x(s_{k+1}))}}$$

c'est-à-dire par :

$$e^{-(s-s_k)(Q_{\text{séq}}(\tilde{x}(s_k))-Q_{\text{séq}}(x(s_k)))}e^{-(s_{k+1}-s)(Q_{\text{séq}}(\tilde{x}(s_{k+1}))-Q_{\text{séq}}(x(s_{k+1})))}\frac{Q_{\text{séq}}(\tilde{x}(s_k),\tilde{x}(s_{k+1}))}{Q_{\text{séq}}(x(s_k),x(s_{k+1}))}.$$

Par le lemme 6.1.4, on sait que les termes sous formes exponentielles ne dépendent pas du choix de $x_{1:i-3}$ et $x_{i+3:m}$. On montre maintenant que le quotient $Q := \frac{Q_{\text{séq}}(\tilde{x}(s_k),\tilde{x}(s_{k+1}))}{Q_{\text{séq}}(x(s_k),x(s_{k+1}))}$ ne dépend pas non plus du choix de $x_{1:i-3}$ et $x_{i+3:m}$. On distingue les cas suivant le site où la substitution a eu lieu :

- si la substitution a lieu sur un site l parmi $1 : i-2$ ou $i+2 : m$, alors :

  $$Q_{\text{séq}}(x(s_k),x(s_{k+1})) = \boldsymbol{Q}_{x_{l-1}(s_k),x_{l+1}(s_k)}(x_l(s_k),x_l(s_{k+1}))$$

  et comme $i \notin \{l-1, l, l+1\}$, on obtient $Q_{\text{séq}}(x(s_k),x(s_{k+1})) = Q_{\text{séq}}(\tilde{x}(s_k),\tilde{x}(s_{k+1}))$. On en déduit que $Q = 1$.
- si la substitution a lieu sur le site $i-1$ (respectivement i, $i+1$), alors le numérateur et le dénominateur du quotient Q ne dépendent pas du choix de $x_{1:i-3}$ et $x_{i+3:m}$ puisque pour $\boldsymbol{x} \in \{x, \tilde{x}\}$, on a respectivement :

  $$Q_{\text{séq}}(\boldsymbol{x}(s_k),\boldsymbol{x}(s_{k+1})) = \boldsymbol{Q}_{\boldsymbol{x}_{i-2}(s_k),\boldsymbol{x}_i(s_k)}(\boldsymbol{x}_{i-1}(s_k),\boldsymbol{x}_{i-1}(s_{k+1})),$$
  $$Q_{\text{séq}}(\boldsymbol{x}(s_k),\boldsymbol{x}(s_{k+1})) = \boldsymbol{Q}_{\boldsymbol{x}_{i-1}(s_k),\boldsymbol{x}_{i+1}(s_k)}(\boldsymbol{x}_i(s_k),\boldsymbol{x}_i(s_{k+1})),$$
  $$Q_{\text{séq}}(\boldsymbol{x}(s_k),\boldsymbol{x}(s_{k+1})) = \boldsymbol{Q}_{\boldsymbol{x}_i(s_k),\boldsymbol{x}_{i+2}(s_k)}(\boldsymbol{x}_{i+1}(s_k),\boldsymbol{x}_{i+1}(s_{k+1})).$$

Par suite, on obtient le résultat globalement : la densité conditionnelle $f(x_i|x_{1:i-1}, x_{i+1:m})$ ne dépend pas du choix de $x_{1:i-3}$ et $x_{i+3:m}$, ce qui conclut la preuve de la proposition.

□

6.1.2 Structure de champ markovien d'ordre un

Pour un modèle d'évolution avec dépendance inclus dans la classe RN95+YpR, le théorème 6.1.3 peut être renforcé pour obtenir un champ markovien d'ordre un en les sites. Pour éviter de gérer les conditions de bords, on considère une séquence évolutive Φ-encodée $((\Phi X)_i)_{i\in[\![1,m]\!]}$ (voir la définition 3.1.2). On utilise alors la notation suivante, adaptée de la notation 1.2.1.

Notation 6.1.5. *On note $\Phi X = (\Phi X(t))_{t\in[0,T]} = ((\Phi X)_i)_{i\in[\![1,m]\!]}$ une évolution Φ-encodée issue d'un modèle d'évolution évoluant du temps initial au temps final T sur m sites. À chaque date $t \in [0,T]$, on écrit :*

$$\Phi X(t) = (\rho(X_1)(t), X_2(t), \ldots, X_{m-1}(t), \eta(X_m)(t)).$$

Le fait de choisir le modèle dans la classe RN95+YpR correspond à choisir la matrice de taux de sauts $Q_{\text{séq}}$ de la section 6.1.1 comme associée aux matrices de taux de sauts $Q_{g,d}$ de la définition 1.2.5. On pose donc dans cette section : $\boldsymbol{Q}_{g,d} = Q_{g,d}$ (pour $(g,d) \in \mathcal{A}^2$).

On peut maintenant énoncer le théorème structurel spatial de champ markovien d'ordre un :

Théorème 6.1.6. *La séquence évolutive Φ-encodée $((\Phi X)_i)_{i\in[\![1,m]\!]}$ est un champ markovien d'ordre un (c'est-à-dire que l'évolution d'un site i conditionnellement aux autres sites ne dépend que des sites $i-1$ et $i+1$ immédiatement voisins à i).*

Pour montrer ce résultat, on démontre tout d'abord le lemme suivant :

Lemme 6.1.7. *On fixe un site $i \in [\![1,m]\!]$, on considère une séquence $x(t) = (x_1, \dots x_m) \in \mathcal{A}^m$ et $N \in \mathcal{A}$.*

Le taux de saut $Q_{x_{i-2}(t),x_i(t)}(x_{i-1}(t), N)$ ne dépend pas :
- *de $x_i(t)$ si $x_{i-1}(t) \in \{A, G\}$,*
- *de $x_{i-2}(t)$ si $x_{i-1}(t) \in \{C, T\}$.*

De même, le taux de saut $Q_{x_i(t),x_{i+2}(t)}(x_{i+1}(t), N)$ ne dépend pas :
- *de $x_i(t)$ si $x_{i+1}(t) \in \{C, T\}$,*
- *de $x_{i+2}(t)$ si $x_{i+1}(t) \in \{A, G\}$.*

Le lemme 6.1.4 se réexprime alors de la façon suivante : pour deux évolutions $x = (x_1, \dots, x_m)$ et $\tilde{x} = (x_1, \dots x_{i-1}, \tilde{x}_i, x_{i+1}, \dots, x_m)$ ne différant qu'au site i et pour tout instant $t \in [0, T]$, la différence $Q_{séq}(x(t)) - Q_{séq}(\tilde{x}(t))$ ne dépend que des nucléotides $x_{i-1}(t), x_i(t), \tilde{x}_i(t)$ et $x_{i+1}(t)$.

Démonstration. D'après la forme des matrices $Q_{g,d}$ (voir définition 1.2.5), la première partie est vérifiée.

On reprend ensuite le terme D_i du lemme 6.1.4 et on peut alors écrire :

$$D_i(x_{i-2:i+2}(t)) = \sum_{j=i-1}^{i+1} \sum_{N\neq x_j(t)} Q_{x_{j-1}(t),x_{j+1}(t)}(x_j(t), N)$$

comme la somme de deux termes $E_i(x_{i-2:i-1}(t), x_{i+1:i+2}(t))$ et $F_i(x_{i-1:i+1}(t))$, ce qui conclut la preuve.

Par exemple, si $x_{i-1}(t) = C$ et $x_{i+1}(t) = C$, on choisit :

$$E_i(x_{i-2:i-1}(t), x_{i+1:i+2}(t)) = \sum_{j=i+1}^{i+1} \sum_{N\neq x_j(t)} Q_{x_{j-1}(t),x_{j+1}(t)}(x_j(t), N)$$

et

$$F_i(x_{i-1:i+1}(t)) = \sum_{j=i-1}^{i} \sum_{N\neq x_j(t)} Q_{x_{j-1}(t),x_{j+1}(t)}(x_j(t), N).$$

□

À l'aide de ce lemme, on montre le théorème 6.1.6.

Démonstration. On reprend la preuve du théorème 6.1.3.

On choisit $x = (x_1, \dots, x_m)$ et $\tilde{x} = (x_1, \dots x_{i-1}, \tilde{x}_i, x_{i+1}, \dots, x_m)$ deux évolutions ne différant qu'au site i, et on cherche à exprimer le rapport de densités $\frac{f(\tilde{x})}{f(x)}$. Comme dans la preuve du théorème 6.1.3, on choisit des instants $s_1 < \dots < s_L$ de telle sorte que sur chaque intervalle $[s_k, s_{k+1}]$, une seule substitution a eu lieu dans $x_1, \dots x_m, \tilde{x}_i$.

On choisit un intervalle $[s_k, s_{k+1}]$, on note s l'instant de la substitution. Le quotient de densité s'exprime alors d'après la proposition 1.5.4 par :

$$e^{-(s-s_k)(Q_{\text{séq}}(\tilde{x}(s_k))-Q_{\text{séq}}(x(s_k)))} e^{-(s_{k+1}-s)(Q_{\text{séq}}(\tilde{x}(s_{k+1}))-Q_{\text{séq}}(x(s_{k+1})))} \frac{Q_{\text{séq}}(\tilde{x}(s_k), \tilde{x}(s_{k+1}))}{Q_{\text{séq}}(x(s_k), x(s_{k+1}))}.$$

Par le lemme 6.1.7, on sait que les termes sous formes exponentielles ne dépendent pas du choix de $x_{1:i-2}$ et $x_{i+2:m}$. On montre maintenant que le quotient $Q := \frac{Q_{\text{séq}}(\tilde{x}(s_k),\tilde{x}(s_{k+1}))}{Q_{\text{séq}}(x(s_k),x(s_{k+1}))}$ ne dépend pas non plus du choix de $x_{1:i-2}$ et $x_{i+2:m}$. On distingue encore les cas suivant le site où la substitution a eu lieu :

- si la substitution a lieu sur un site l parmi $1 : i-2$ ou $i+2 : m$, alors le quotient Q est égal à 1.
- si la substitution a lieu sur le site i, alors le numérateur et le dénominateur du quotient Q ne dépendent pas du choix de $x_{1:i-2}$ et $x_{i+2:m}$ puisque pour $\boldsymbol{x} \in \{x, \tilde{x}\}$:

$$Q_{\text{séq}}(\boldsymbol{x}(s_k), \boldsymbol{x}(s_{k+1})) = Q_{\boldsymbol{x}_{i-1}(s_k),\boldsymbol{x}_{i+1}(s_k)}(\boldsymbol{x}_i(s_k), \boldsymbol{x}_i(s_{k+1})).$$

- si la substitution a lieu sur le site $i-1$, on écrit pour $\boldsymbol{x} \in \{x, \tilde{x}\}$:

$$Q_{\text{séq}}(\boldsymbol{x}(s_k), \boldsymbol{x}(s_{k+1})) = Q_{\boldsymbol{x}_{i-2}(s_k),\boldsymbol{x}_i(s_k)}(\boldsymbol{x}_{i-1}(s_k), \boldsymbol{x}_{i-1}(s_{k+1})).$$

  On a alors :
  - si $x_{i-1}(t) \in \{C, T\}$, par le lemme 6.1.7, $Q_{\text{séq}}(\boldsymbol{x}(s_k), \boldsymbol{x}(s_{k+1}))$ ne dépend pas de $x_{1:i-2}$ ni de $x_{i+2:m}$.
  - si $x_{i-1}(t) \in \{A, G\}$, par le lemme 6.1.7, $Q_{\text{séq}}(\boldsymbol{x}(s_k), \boldsymbol{x}(s_{k+1}))$ ne dépend pas de $\boldsymbol{x}_i$ donc on obtient $Q_{\text{séq}}(x(s_k), x(s_{k+1})) = Q_{\text{séq}}(\tilde{x}(s_k), \tilde{x}(s_{k+1}))$ et le quotient Q vaut 1.
- si la substitution a lieu sur le site $i+1$, on raisonne de même pour obtenir que le quotient ne Q ne dépend pas du choix de $x_{1:i-2}$ et $x_{i+2:m}$.

Par suite, on obtient le résultat globalement : la densité conditionnelle $f(x_i|x_{1:i-1}, x_{i+1:m})$ ne dépend pas du choix de $x_{1:i-2}$ et $x_{i+2:m}$, ce qui conclut la preuve de la proposition. □

Remarque 6.1.8. *Le théorème 6.1.6 ne contredit pas la représentation schématique des dépendances de la figure 2.1. En effet, on fait évoluer ici uniquement le site i conditionnellement aux sites voisins encodés. Ces sites voisins sont fixés et on ne les fait donc pas évoluer. Ainsi pour les sites voisins, on ne cherche pas l'évolution ρ-encodée (resp. η-encodée) du site $i-1$ (resp. $i+1$) conditionnée par les feuilles (non encodées), mais simplement la probabilité d'une évolution fixée encodée dans ρ (resp. η).*

Nous allons maintenant expliciter ce théorème, pour donner un moyen d'effectuer l'évolution d'un site conditionnellement aux sites voisins.

6.1.3 Écriture de l'évolution

On cherche à décrire l'évolution en un site $i \in [\![1, m]\!]$ fixé et à l'instant $t \in [0, T]$, connaissant en tout temps l'évolution aux sites $i-1$ et $i+1$. Cela est possible car on sait que la loi de l'évolution au site i conditionnellement aux voisins est markovienne inhomogène en temps.

On suppose être en $x \in \mathcal{A}$ au temps t et on souhaite connaître le taux de substitution instantané vers $x' \in \mathcal{A}$ en cet instant. Pour écrire explicitement le taux de substitution instantané au temps t, nous allons utiliser les éléments suivants.

Définition 6.1.9. *Pour un temps $t \in [0, T]$, on définit $t_0 = t$, $t_L = T$ et $(t_l)_{l\in[\![1,L-1]\!]}$ la succession des instants de changements affectant $\rho(x_{i-1})$ ou $\eta(x_{i+1})$ à partir du temps t_0. On écrit alors pour $l \in [\![1, L]\!]$:*

$$\Delta_l = t_l - t_{l-1}.$$

Sur chaque intervalle $[t_{l-1}, t_l[$, $\rho(x_{i-1})$ et $\eta(x_{i+1})$ sont donc constants et valent respectivement $\rho(x_{i-1}(t_{l-1}))$ et $\eta(x_{i+1}(t_{l-1}))$.

Définition 6.1.10. *Pour $g, d \in \mathcal{A}$, on définit la matrice $U_{g,d}$ de taille 4×4 définie pour $x, x' \in \mathcal{A}$ par :*

$$U_{g,d}(x,x') = \begin{cases} Q_{,x}(g,g) + Q_{g,d}(x,x) + Q_{x,}(d,d) & \text{si } x = x' \\ Q_{g,d}(x,x') & \text{sinon.} \end{cases}$$

Les matrices $U_{g,d}$ ne sont pas des matrices de taux de sauts puisque la somme des termes de chaque ligne est strictement négative. On verra qu'elles permettent néanmoins de quantifier l'évolution au site i lorsque l'état des voisins immédiats demeure inchangé. De telles matrices, c'est-à-dire vérifiant que la somme des termes de chaque ligne est strictement négative sont étudiées dans [2] sous le nom *dishonest Q-function.*

On définit alors la fonction suivante qui permettra de pondérer le taux de substitution instantané en fonction de l'évolution sur les sites $i-1$ et $i+1$.

Définition 6.1.11. *Pour $t \in [0,T]$, on considère les durées $(\Delta_l)_{l \in [\![1,L]\!]}$ associées issues de la définition 6.1.9. On définit la matrice $H(t)$ de taille 4×4 par :*

$$H(t) = \prod_{l=1}^{L} \exp\left(\Delta_l U_{x_{i-1}(t_{l-1}), x_{i+1}(t_{l-1})}\right).$$

On peut maintenant énoncer la proposition suivante.

Proposition 6.1.12. *Le taux de saut de x vers x' au site i et à l'instant t, connaissant $x_{i-1}(0 \to T)$ et $x_{i+1}(0 \to T)$, est donné pour tous $x \neq x'$ par :*

$$\frac{\sum_{x_T \in \mathcal{A}} H(t)(x', x_T)}{\sum_{x_T \in \mathcal{A}} H(t)(x, x_T)} Q_{x_{i-1}(t), x_{i+1}(t)}(x, x').$$

Remarque 6.1.13. *Le terme $Q_{x_{i-1}(t), x_{i+1}(t)}(x, x')$ correspond au taux de saut lorsque l'on connaît uniquement aux temps précédant t les évolutions x_{i-1} et x_{i+1}. Ainsi, à ce terme il faut rajouter le terme correctif correspondant au premier terme.*

Avant de passer à la preuve de cette proposition, voyons à travers l'exemple 6.1.14 que ce taux de substitution instantané dépend entièrement des évolutions connues $x_{i-1}(t \to T)$ et $x_{i+1}(t \to T)$, et non uniquement des nucléotides au temps considéré $x_{i-1}(t)$ et $x_{i+1}(t)$.

Exemple 6.1.14. *On considère le modèle suivant (les taux non indiqués sont nuls) pour $\varepsilon \ll 1$:*

$$v_A = 1, r_{CA \to CG} = 10 \text{ et } w_x = v_{x'} = \varepsilon \text{ pour } x \in \mathcal{A}, x' \in \{C, G, T\}.$$

On suppose que $x_{i-1} \equiv A$ et que l'évolution x_{i+1} est donnée par :

t	*i+1*
0	*C*
0.89	*A*
0.90	*G*
1	*G*

On suppose qu'au temps initial, le nucléotide au site i est C et on veut connaître le taux de substitution de C vers A à cet instant quand ε tend vers 0.

On remarque tout d'abord qu'avec $x_{i-1} \equiv A$ et x_{i+1} quelconque, les deux types d'évolution au site i débutant en C et dont la densité ne tend pas vers 0 (quand ε tend vers 0) sont :

- *$x_i \equiv C$ sur tout l'intervalle $[0,1]$ et*
- *$x_i(t) = C$ pour $t < t_0$ et $x_i(t) = A$ pour $t \geq t_0$, avec $t_0 \in]0,1[$.*

De plus, pour que la densité de l'évolution au site $i+1$ ne tende pas vers 0, il faut que le marqueur associé à la substitution survenue à l'instant 0.90 provienne d'un marqueur associé au taux $r_{CA\to CG}$. Cela n'est possible que si le nucléotide au site i à cet instant est C.

Ainsi, la densité globale de l'évolution (x_{i-1}, x_i, x_{i+1}) ne tend pas vers 0 uniquement lorsque $x_i \equiv C$ sur tout l'intervalle $[0,1]$.

Le taux de substitution de C vers A à l'instant 0 tend donc vers 0. Par contre, si on connaît uniquement $x_{i-1}(t)$ et $x_{i+1}(t)$ à l'instant $t = 0$, le taux de substitution de C vers A à l'instant 0 est $v_A = 1$.

On démontre maintenant la proposition 6.1.12.

Démonstration. On considère $x \neq x'$ deux nucléotides et des instants $0 < s < t < T$. On abrège $X_i(s) = x$ en x et $X_i(t) = x'$ en x'. On suppose de plus qu'à l'instant choisi, aucune substitution n'a eu lieu aux sites $i-1$ et $i+1$. On cherche à calculer la probabilité conditionnelle : $P(x' \mid x_{i-1}, x, x_{i+1})$. On écrit :

$$\begin{aligned} P(x' \mid x_{i-1}, x, x_{i+1}) =& f(x' \mid x_{i-1}(s \to T), x, x_{i+1}(s \to T)) \\ =& \frac{f(x_{i-1}(s \to T), x, x', x_{i+1}(s \to T))}{f(x_{i-1}(s \to T), x, x_{i+1}(s \to T))}. \end{aligned}$$

Par conditionnement au numérateur et au dénominateur et en utilisant le caractère markovien dans le temps de l'évolution aux instants t et s, $P(x' \mid x_{i-1}, x, x_{i+1})$ devient :

$$\frac{f(x_{i-1}(t\to T), x_{i+1}(t\to T) \mid x_{i-1}(t), x', x_{i+1}(t))}{f(x_{i-1}(s\to T), x_{i+1}(s\to T) \mid x_{i-1}(s), x, x_{i+1}(s))} f(x_{i-1}(s \to t), x', x_{i+1}(s \to t) | x_{i-1}(s), x, x_{i+1}(s)).$$

En écrivant pour tous $t < t'$ et $x, x' \in \mathcal{A}$:

$$h(t \to t')(x, x') = f(x_{i-1}(t \to t'), X_i(t') = x', x_{i+1}(t \to t') \mid x_{i-1}(t), X_i(t) = x, x_{i+1}(t)),$$

on a alors :

$$P(x' \mid x_{i-1}, x, x_{i+1}) = \frac{\sum_{x_T \in \mathcal{A}} h(t \to T)(x', x_T)}{\sum_{x_T \in \mathcal{A}} h(s \to T)(x, x_T)} h(s \to t)(x, x'). \tag{6.1}$$

Il reste à étudier la fonction h. On remarque d'abord, en utilisant le fait que l'évolution est markovienne dans le temps, que pour tout $t' \in]t, t''[$,

$$h(t \to t'')(x, x'') = \sum_{x' \in \mathcal{A}} h(t \to t')(x, x') h(t' \to t'')(x', x'').$$

On découpe alors l'intervalle $[t, T[$ comme énoncé dans la définition 6.1.9. Prenons un morceau $[t_l, t_{l+1}[$. Sur ce morceau, il existe $g \in \{R, C, T\}$ et $d \in \{A, G, Y\}$ tels que :

$\rho(x_{i-1}) \equiv g$ et $\eta(x_{i+1}) \equiv d$. Pour $\varepsilon > 0$, le taux de transition instantané $u^{\varepsilon}_{g,d}$ de x vers x' pour un instant de cet intervalle est donné par :

$$\begin{aligned} u^{\varepsilon}_{g,d}(x,x') =& q^{\varepsilon}_{\cdot,x}(g,g) q^{\varepsilon}_{g,d}(x,x') q^{\varepsilon}_{x,\cdot}(d,d) \\ =& [1 + Q_{\cdot,x}(g,g)\varepsilon] q^{\varepsilon}_{g,d}(x,x') [1 + Q_{x,\cdot}(d,d)\varepsilon] \\ =& \begin{cases} 1 + (Q_{\cdot,x}(g,g) + Q_{g,d}(x,x) + Q_{x,\cdot}(d,d))\,\varepsilon + o(\varepsilon) & \text{si } x = x' \\ Q_{g,d}(x,x')\varepsilon + o(\varepsilon) & \text{sinon.} \end{cases} \end{aligned}$$

Ainsi $u^{\varepsilon}_{g,d}$ est associé à la matrice $U_{g,d}$ de la définition 6.1.10, et on écrit :

$$h(t_l \to t_{l+1}) = \exp\left((t_{l+1} - t_l) U_{g,d}\right).$$

On obtient donc $h(t \to T) = H(t)$ puis par continuité de H en t :

$$P(x' \mid x_{i-1}, x, x_{i+1}) = \frac{\sum_{x_T \in \mathcal{A}} H(t)(x', x_T)}{\sum_{x_T \in \mathcal{A}} H(t)(x, x_T)} Q_{x_{i-1}(t), x_{i+1}(t)}(x, x')(t - s) + o(t - s),$$

ce qui conclut la proposition. □

Conditionnement par le site final. En général, on conditionne aussi par le nucléotide noté x_T au site i associé aux feuilles. La proposition 6.1.12 s'adapte alors facilement. On observe que la forme de la matrice s'exprime comme une h-transformée de Doob d'une chaîne de Markov, ici associée au conditionnement de l'évolution par le site final (voir [75] section 17.6 et [119] Chapitre III, théorème 49.3).

Proposition 6.1.15. *Le taux de saut de x vers x' au site i et à l'instant t, connaissant $x_{i-1}(0 \to T)$, $x_i(T) = x_T$ et $x_{i+1}(0 \to T)$, est donné pour tous $x \neq x'$ par :*

$$\frac{H(t)(x', x_T)}{H(t)(x, x_T)} Q_{x_{i-1}(t), x_{i+1}(t)}(x, x').$$

On en déduit en chaque instant la matrice de taux de sauts instantanée associée à l'évolution au site i conditionnellement aux évolutions x_{i-1} et x_{i+1} et au nucléotide final $x_i(T)$.

Définition 6.1.16. *On suppose fixés les évolutions x_{i-1} et x_{i+1} ainsi que le nucléotide $x_i(T) = x_T$.*

Pour tout $t \in [0, T]$, on définit $\tilde{Q}(t)$ la matrice de taux de sauts sur $\mathcal{A}$ définie pour $x \neq x' \in \mathcal{A}$ par :

$$\tilde{Q}(t)(x, x') = \frac{H(t)(x', x_T)}{H(t)(x, x_T)} Q_{x_{i-1}(t), x_{i+1}(t)}(x, x')$$

et telle que la somme sur chacun des lignes de la matrice soit nulle.

6.1.4 Fonction de survie

On déduit de la section 6.1.3 l'écriture de la fonction de survie pour l'évolution d'un site conditionnellement à l'évolution du site précédent et suivant.

Pour tout instant t, on reprend la matrice de taux de sauts $\tilde{Q}(t)$ (définition 6.1.16) associée à l'évolution au site i conditionnellement aux évolutions x_{i-1} et x_{i+1} et au nucléotide final $x_i(T)$.

On cherche à calculer la fonction de survie associée à $\tilde{Q}$ à partir d'un instant t_0, définie de la façon suivante :

Définition 6.1.17. *La fonction de survie à partir de l'instant t_0 et associée à la matrice de taux de sauts $\tilde{Q}$ est notée $\bar{F}_{t_0}$. Elle vérifie, pour tout $t > t_0$ et tout $x \in \mathcal{A}$:*

$$\bar{F}_{t_0}(t)(x) = P(X_i(t_0 \to t) \equiv x | X_i(t_0) = x, X_i(T) = x_T, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}).$$

Puisque $\tilde{Q}$ est une matrice de taux de sauts, on sait que $\bar{F}_{t_0}$ vérifie en outre la relation (avec $\tilde{Q}(t_0)(.) = -\tilde{Q}(t_0)(.,.) > 0$) :

$$\bar{F}'_{t_0}(t)(x) = \bar{F}_{t_0}(t)(x)\tilde{Q}(t)(x,x).$$

La proposition suivante permet d'expliciter $\bar{F}_{t_0}$ jusqu'au prochain instant t_1 affectant $\rho(x_{i-1})$ ou $\eta(x_{i+1})$ (voir la définition 6.1.9).

Proposition 6.1.18. *Pour $t \in [t_0, t_1[$ (où t_1 est défini selon la définition 6.1.9) et $x \in \mathcal{A}$, la valeur de la fonction de survie $\bar{F}_{t_0}(t)(x)$ est donnée par :*

$$\exp\left((t - t_0)U_{x_{i-1}(t_0),x_{i+1}(t_0)}(x,x)\right) \frac{H(t)(x, x_T)}{H(t_0)(x, x_T)}.$$

Démonstration. Sur l'intervalle $[t_0, t_1[$, on sait qu'aucun changement n'a affecté $\rho(x_{i-1})$ ou $\eta(x_{i+1})$ donc pour tout $t \in [t_0, t_1[$, on a $\tilde{Q}(t) = \tilde{Q}(t_0)$.

Ensuite, pour $\varepsilon > 0$, on découpe l'intervalle $[t_0, t[$ en une partition dont chaque morceau est de longueur plus petit que ε. On note $t_0 = s_0 < s_1 < \ldots < s_L = t$ les instants associés. On écrit puisque l'évolution est markovienne dans le temps :

$$\bar{F}_{t_0}(t)(x) = \prod_{k=1}^{L} \bar{F}_{s_{k-1}}(s_k)(x). \tag{6.2}$$

On reprend les notations de la preuve de la proposition 6.1.12, en particulier l'équation (6.1). Pour tout $k \in [\![1, L]\!]$, comme la probabilité que plus de deux sauts se produisent au site i sur l'intervalle $[s_k, s_{k+1}[$ est négligeable devant $s_{k+1} - s_k \leq \varepsilon$, on écrit :

$$\begin{aligned}
\bar{F}_{s_{k-1}}(s_k)(x) =& P(X_i(s_k) = x \mid x_{i-1}, X_i(s_{k-1}) = x, X_i(T) = x_T, x_{i+1}) + o(\varepsilon) && (6.3)\\
=& \frac{h(s_k \to T)(x, x_T)}{h(s_{k-1} \to T)(x, x_T)} h(s_{k-1} \to s_k)(x,x) + o(\varepsilon). && (6.4)
\end{aligned}$$

En effectuant le produit, l'équation (6.2) devient (puisque le premier terme dans l'équation (6.4) est $O(1)$) :

$$\frac{h(s_L \to T)(x, x_T)}{h(s_0 \to T)(x, x_T)} \prod_{k=1}^{L} h(s_{k-1} \to s_k)(x,x) + o(\varepsilon). \tag{6.5}$$

Il reste à étudier le terme $S := \prod_{k=1}^{L} h(s_{k-1} \to s_k)(x,x)$. En passant au logarithme, on obtient :

$$\begin{aligned}
\log S =& \sum_{k=1}^{L} \log(1 + U_{x_{i-1}(t_0),x_{i+1}(t_0)}(x,x)(s_k - s_{k-1}) + o(\varepsilon))\\
=& \sum_{k=1}^{L} \left(U_{x_{i-1}(t_0),x_{i+1}(t_0)}(x,x)(s_k - s_{k-1}) + o(\varepsilon)\right)\\
=& U_{x_{i-1}(t_0),x_{i+1}(t_0)}(x,x)(s_L - s_0) + o(1).
\end{aligned}$$

Enfin, en utilisant que $s_0 = t_0$ et $s_L = t$, puis en faisant tendre ε vers 0, on en déduit à partir de l'équation (6.5) :

$$\bar{F}_{t_0}(t)(x) = \exp\left((t-t_0)U_{x_{i-1}(t_0),x_{i+1}(t_0)}(x,x)\right)\frac{H(t)(x,x_T)}{H(t_0)(x,x_T)}.$$

□

6.1.5 Structure associée de chaîne de Markov

La structure spatiale de champ markovien de l'évolution $(X_i)_{i\in[\![1,m]\!]}$ issue d'un modèle RN95+YpR induit une structure spatiale de chaîne de Markov d'après le théorème de Hammersley-Clifford [16] (la condition de positivité est ici clairement vérifiée). On peut le redémontrer ici directement dans notre cas, c'est-à-dire pour un champ markovien sur $[\![1,m]\!]$. On considère une densité f par rapport à une mesure μ et associée à une évolution $(X_{1:m})$.

On définit tout d'abord la condition de positivité, vérifiée ici d'après la proposition 1.5.4 exprimant la densité des évolutions.

Définition 6.1.19. *On dit que f vérifie la condition de positivité si pour tous $x_1,\dots,x_m$ tels que $f(x_k)>0$ pour tout $k\in[\![1,m]\!]$, alors $f(x_{1:m})>0$.*

L'égalité suivante correspond à l'équation (2.2) de [16], dont on reprend la preuve.

Lemme 6.1.20. *On suppose que la densité f vérifie la condition de positivité. Soit $x_{1:m}$ et $y_{1:m}$ deux évolutions, avec $f(y_{1:m})>0$. On a l'égalité suivante :*

$$\frac{f(y_{1:m})}{f(x_{1:m})} = \prod_{k=1}^{m}\frac{f(y_k|y_{1:k-1},x_{k+1:m})}{f(x_k|y_{1:k-1},x_{k+1:m})}.$$

Démonstration. On écrit tout d'abord (la condition de positivité assure que le dénominateur ne s'annule pas) :

$$f(y_{1:m}) = \frac{f(y_m|y_{1:m-1})}{f(x_m|y_{1:m-1})}f(y_{1:m-1},x_m).$$

Le dernier terme $f(y_{1:m-1},x_m)$ s'écrit ensuite

$$f(y_{1:m-1},x_m) = \frac{f(y_{m-1}|y_{1:m-2},x_m)}{f(x_{m-1}|y_{1:m-2},x_m)}f(y_{1:m-2},x_{m-1:m}).$$

Par récurrence sur les sites, on obtient alors après m étapes :

$$f(y_{1:m}) = \prod_{k=1}^{m}\frac{f(y_k|y_{1:k-1},x_{k+1:m})}{f(x_k|y_{1:k-1},x_{k+1:m})}f(x_{1:m}),$$

d'où la conclusion. □

Le lemme 6.1.20 permet d'en déduire que lorsque $(X_{1:m})$ est un champ markovien (et vérifie la condition de positivité), alors pour tout site i l'évolution sur $[\![i,m]\!]$ ne dépend de l'évolution sur $[\![1,i-1]\!]$ qu'à travers $i-1$.

Lemme 6.1.21. *Soit $(X_{1:m})$ un champ markovien associé à une densité f vérifiant la condition de positivité. Soit $x_{1:m}$ une évolution vérifiant $f(x_{1:i-1}) > 0$. Alors on a l'égalité suivante :*

$$f(x_{i:m}|x_{1:i-1}) = f(x_{i:m}|x_{i-1}).$$

Démonstration. Pour tout $i \in [\![1, m]\!]$, on écrit le premier terme de l'égalité à vérifier sous la forme suivante :

$$f(x_{i:m}|x_{1:i-1}) = \frac{f(x_{1:m})}{\int_{\tilde{x}_{i:m}} f(x_{1:i-1}, \tilde{x}_{i:m}) d\mu(\tilde{x}_{i:m})}.$$

On applique ensuite pour chaque terme $f(x_{1:i-1}, \tilde{x}_{i:m})$ le lemme 6.1.20 avec $y_{1:m} = (x_{1:i-1}, \tilde{x}_{i:m})$. On obtient alors :

$$\frac{f(x_{1:i-1}, \tilde{x}_{i:m})}{f(x_{1:m})} = \prod_{k=1}^{i-1} \frac{f(x_k|x_{1:k-1}, y_{k+1:m})}{f(x_k|x_{1:k-1}, y_{k+1:m})} \prod_{k=i}^{m} \frac{f(x_k|x_{1:k-1}, \tilde{x}_{k+1:m})}{f(\tilde{x}_k|x_{1:k-1}, \tilde{x}_{k+1:m})}.$$

Le premier produit vaut 1 et par le fait que l'on a un champ markovien d'ordre 1, on en déduit :

$$\frac{f(x_{1:i-1}, \tilde{x}_{i:m})}{f(x_{1:m})} = \prod_{k=i}^{m} \frac{f(x_k|x_{k-1}, \tilde{x}_{k+1})}{f(\tilde{x}_k|x_{k-1}, \tilde{x}_{k+1})},$$

ce dernier terme ne dépendant de $x_{1:i-1}$ qu'à travers x_{i-1}. □

On en conclut que sous la condition de positivité, un champ markovien sur $[\![1, m]\!]$ est une chaîne de Markov (d'ordre 1).

Proposition 6.1.22. *Soit $(X_{1:m})$ un champ markovien associé à une densité f vérifiant la condition de positivité. Alors $(X_{1:m})$ est une chaîne de Markov.*

Démonstration. Pour une évolution $x_{1:m}$ de densité positive, on a pour tout $i \in [\![1, m]\!]$:

$$f(x_i|x_{1:i-1}) = \int_{x_{i+1:m}} f(x_{i:m}|x_{1:i-1}) d\mu(x_{i+1:m}).$$

Par le lemme 6.1.21, on obtient le résultat voulu :

$$f(x_i|x_{1:i-1}) = \int_{x_{i+1:m}} f(x_{i:m}|x_{i-1}) d\mu(x_{i+1:m}) = f(x_i|x_{i-1}).$$

□

La forme explicite des transitions associée à la chaîne de Markov ainsi obtenue n'est pas utilisable directement puisque d'après le lemme 6.1.21, elle fait intervenir l'ensemble des évolutions possibles pour les sites à droite du site considéré.

6.2 Structure de chaîne de Markov explicite

Nous avons exhibé dans la section précédente une structure spatiale de champ markovien de l'évolution dans le modèle RN95+YpR. Elle induit une structure spatiale de chaîne de Markov, mais dont la substitution vers chaque site fait intervenir l'ensemble des évolutions possibles des sites suivants le site considéré.

Nous construisons dans cette section une structure spatiale markovienne de l'évolution dont les transitions spatiales s'expriment explicitement sans avoir besoin de considérer les sites futurs. Pour cela, nous allons non pas regarder l'évolution spatiale site par site mais encoder chaque site de deux façons différentes et considérer l'évolution à cheval sur deux sites consécutifs encodés sous la forme (ρ, η) (section 6.2.1). On vérifie en outre que cette construction permet de reconstituer entièrement l'évolution initiale site par site.

On établit une propriété de chaîne de Markov de cette évolution encodée, dont on explicite le comportement des substitutions en chaque site conditionnellement au site précédent, sans (section 6.2.2, avec les preuves associées sections 6.2.3 et 6.2.4) ou avec (section 6.2.5) conditionnement par la séquence associée au temps actuel. Comme pour la structure de champ markovien, on en déduit l'écriture explicite de la fonction de survie (section 6.2.6).

L'écriture de cette fonction est utile pour simuler l'évolution encodée en tout site conditionnellement au site précédent, et permet d'utiliser des méthodes particulaires pour la simulation des évolutions et le calcul consistant de la vraisemblance d'une séquence (voir chapitre 8).

6.2.1 Évolution en termes d'encodages (ρ, η)

Au lieu de regarder l'évolution site par site, nous considérons l'évolution encodée (ρ, η) par (ρ, η) (voir remarque 3.1.5), c'est-à-dire uniquement des séquences Φ-encodées de longueur 2. On reprend alors la définition 3.2.3 rappelée ci-après.

Définition 6.2.1. *Pour chaque entier $i \in [\![1, m-1]\!]$, on appelle Z_i l'évolution (ρ, η) aux sites $(i, i+1)$ définie par :*

$$Z_i := (\rho_i, \eta_{i+1}).$$

On rappelle que la connaissance pour un site de η_i et de ρ_i permet de reconstituer entièrement l'évolution X_i. On peut alors identifier l'évolution Φ-encodée :

$$\Phi X = (\rho_1, X_2, \ldots, X_{m-1}, \eta_m)$$

avec :

$$(Z_1, \ldots, Z_{m-1}) = (\rho_1, \eta_2, \rho_2, \ldots, \eta_{m-1}, \rho_{m-1}, \eta_m).$$

6.2.2 Structure spatiale de chaîne de Markov

On énonce le théorème structurel de chaîne de Markov de la suite $(Z_i)_{i \in [\![1,m-1]\!]}$ dans le théorème 6.2.2, avant d'expliciter les transitions associées à la chaîne de Markov $(Z_i)_{i \in [\![1,m-1]\!]}$ (théorème 6.2.3 et corollaire 6.2.5). Les preuves de ces théorèmes sont données aux sections 6.2.3 et 6.2.4.

Théorème 6.2.2. *On suppose que la racine $(Z_i(0))_{i \in [\![1,m-1]\!]}$ est fixée. $(Z_i)_{i \in [\![1,m-1]\!]}$ est une chaîne de Markov.*

Théorème 6.2.3. *Soit $t \in]0,T]$ et $i \in [\![1, m-1]\!]$. On suppose que l'on connaît l'évolution en toute date sur les sites $Z_{1:i-1}$ et l'évolution sur l'intervalle de temps $[0,t[$ pour Z_i. Conditionnellement à ces évolutions, la dépendance à gauche dans l'écriture de la probabilité de transition de Z_i à l'instant t se fait seulement à travers les nucléotides π-encodés $\pi_i(t)$ et $\pi_i(t-) := \lim_{\varepsilon \to 0} \pi_i(t-\varepsilon)$. On distingue alors les quatre cas suivants :*

1. *Si $\pi_i(t-) = \pi_i(t) = Y$, l'évolution est donnée par la matrice de taux de sauts suivante (chaque terme de la diagonale est tel que la somme de chaque ligne soit égale à zéro) :*

$$W_Y = \begin{array}{c} \\ CA \\ CG \\ CY \\ TA \\ TG \\ TY \end{array} \begin{array}{c} \begin{array}{cccccc} CA & CG & CY & TA & TG & TY \end{array} \\ \begin{pmatrix} . & w_G + r_{CA\to CG} & v_T + v_C & w_T + r_{CA\to TA} & 0 & 0 \\ w_A + r_{CG\to CA} & . & v_T + v_C & 0 & w_T + r_{CG\to TG} & 0 \\ v_A & v_G & . & 0 & 0 & w_T \\ w_C + r_{TA\to CA} & 0 & 0 & . & w_G + r_{TA\to TG} & v_T + v_C \\ 0 & w_C + r_{TG\to CG} & 0 & w_A + r_{TG\to TA} & . & v_T + v_C \\ 0 & 0 & w_C & v_A & v_G & . \end{pmatrix} \end{array}.$$

2. *Si $\pi_i(t-) = \pi_i(t) = R$, l'évolution est donnée par la matrice de taux de sauts suivante (chaque terme de la diagonale est tel que la somme de chaque ligne soit égale à zéro) :*

$$W_R = \begin{array}{c} \\ RA \\ RG \\ RY \end{array} \begin{array}{c} \begin{array}{ccc} RA & RG & RY \end{array} \\ \begin{pmatrix} . & w_G & v_C + v_T \\ w_A & . & v_C + v_T \\ v_A & v_G & . \end{pmatrix} \end{array}.$$

3. *Si $\pi_i(t-) = Y$ et $\pi_i(t) = R$, la matrice de substitution instantanée de z_i à l'instant t est donnée par :*

$$U_{Y\to R} = \begin{array}{c} \\ CA \\ CG \\ CY \\ TA \\ TG \\ TY \end{array} \begin{array}{c} \begin{array}{ccc} RA & RG & RY \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array}.$$

4. *Si $\pi_i(t-) = R$ et $\pi_i(t) = Y$, la matrice de substitution instantanée de z_i à l'instant t est donnée par :*

$$U_{R\to Y} = \frac{1}{v_C + v_T} \begin{array}{c} \\ RA \\ RG \\ RY \end{array} \begin{array}{c} \begin{array}{cccccc} CA & CG & CY & TA & TG & TY \end{array} \\ \begin{pmatrix} v_C & 0 & 0 & v_T & 0 & 0 \\ 0 & v_C & 0 & 0 & v_T & 0 \\ 0 & 0 & v_C & 0 & 0 & v_T \end{pmatrix} \end{array}.$$

Globalement, on note $\hat{Q}_{t,\pi_i}$ la matrice de taux de saut instantanée associée à l'évolution encodée π_i et à l'instant t.

Définition 6.2.4. *Pour un temps $t_0 \in [0,T]$, on définit $(t_l)_{l \in [\![1,L-1]\!]}$ les instants de changements ordonnés survenant sur $\pi(x_i)$ à partir du temps t_0, et $t_L = T$. On écrit alors pour $l \in [\![1,L]\!]$:*

$$\Delta_l = t_l - t_{l-1}.$$

Corollaire 6.2.5. *On suppose connaître l'évolution de π_i en toute date. On utilise alors les notations de la définition 6.2.4. On note $N_1 = \pi_i(t_0) \in \{R, Y\}$, $N_2 = \{R, Y\} \setminus \{N_1\}$ et pour $l \geq 2$, $N_l = N_{l \bmod (2)}$.*

La matrice de transition (définie dans la propriété 1.5.5) de l'évolution Z_i de t_0 à T conditionnellement à l'évolution de π_i est donnée par :

$$G_{t_0 \to T} = \prod_{l=1}^{L-1} \left(e^{\Delta_l W_{N_l}} U_{N_l \to N_{l+1}} \right) e^{\Delta_L W_{N_L}}.$$

6.2.3 Construction de l'évolution encodée par processus de Poisson

Pour montrer le théorème 6.2.2, on choisit de décrire l'évolution $(Z_{1:i})$ (avec i un site fixé) à partir de la construction par processus de Poisson indépendants de la section 1.2.3 de l'évolution $(X_{1:m})$.

On rappelle que :

$$\mathcal{B}_d = \{CA \to TA,\ CG \to TG,\ TA \to CA,\ TG \to CG\},$$

$$\mathcal{B}_g = \{CA \to CG,\ CG \to CA,\ TA \to TG,\ TG \to TA\}.$$

Pour chaque site k, on partitionne les 16 processus de Poisson indépendants décrits dans la section 1.2.3 en trois ensembles.

D'abord, on considère les processus dont les mouvements associés ne peuvent pas affecter η_k. Il s'agit des mouvements pouvant uniquement effectuer des transitions $C \leftrightarrow T$. L'ensemble des processus considérés sont notés $\Xi_\rho(k)$.

Ensuite, on considère les processus dont les mouvements associés ne peuvent pas affecter ρ_k. Il s'agit des mouvements pouvant uniquement effectuer des transitions $A \leftrightarrow G$. L'ensemble des processus considérés sont notés $\Xi_\eta(k)$.

Enfin, on considère les processus qui peuvent affecter à la fois η_k et ρ_k. Il s'agit uniquement des processus dont les mouvements associés peuvent être des transversions. L'ensemble des processus considérés sont notés $\Xi_0(k)$.

Globalement, on écrit la définition suivante.

Définition 6.2.6. *Pour tout site $k \in [\![1, m]\!]$:*

$\Xi_\rho(k)$ est le regroupement des 6 processus de Poisson suivants :
- *Pour $x \in \{C, T\}$, $\mathcal{W}_k^x$ le processus homogène de Poisson de taux w_x,*
- *Pour $y \in \mathcal{B}_d$, $\mathcal{Q}_k^{y'}$ le processus homogène de Poisson de taux $r_{y'}$.*

$\Xi_\eta(k)$ est le regroupement des 6 processus de Poisson suivants :
- *Pour $x \in \{A, G\}$, $\mathcal{W}_k^x$ le processus homogène de Poisson de taux w_x,*
- *Pour $y \in \mathcal{B}_g$, $\mathcal{R}_k^y$ le processus homogène de Poisson taux r_y.*

$\Xi_0(k)$ est le regroupement des 4 processus de Poisson suivants :
- *Pour $x \in \mathcal{A}$, $\mathcal{V}_k^x$ le processus homogène de Poisson de taux v_x.*

À partir de $\Xi_0(k)$, on définit trois ensembles $\Xi_{0,\pi}(k)$, $\Xi_{0,\eta}(k)$ et $\Xi_{0,\rho}(k)$ par superposition de certains processus.

Définition 6.2.7. *Pour tout site $k \in [\![1, m]\!]$:*

$\Xi_{0,\pi}(k)$ est le regroupement des deux processus de Poisson suivants :

– $\mathcal{V}_k^R := \mathcal{V}_k^A \cup \mathcal{V}_k^G$ *le processus homogène de Poisson de taux* $v_A + v_G$.
– $\mathcal{V}_k^Y := \mathcal{V}_k^C \cup \mathcal{V}_k^T$ *le processus homogène de Poisson de taux* $v_C + v_T$.

Ainsi, l'ensemble $\Xi_{0,\pi}(k)$ *consiste à reprendre* $\Xi_0(k)$ *et à superposer d'une part les processus de Poisson* $\mathcal{V}_k^C$ *et* $\mathcal{V}_k^T$ *et d'autre part les processus de Poisson* $\mathcal{V}_k^A$ *et* $\mathcal{V}_k^G$.

$\Xi_{0,\eta}(k)$ *est le regroupement des deux processus de Poisson suivants :*
– *Pour* $x \in \{A, G\}$, $\mathcal{V}_k^x$ *le processus homogène de Poisson de taux* v_x.
– $\mathcal{V}_k^Y := \mathcal{V}_k^C \cup \mathcal{V}_k^T$ *le processus homogène de Poisson de taux* $v_C + v_T$.

Ainsi, l'ensemble $\Xi_{0,\eta}(k)$ *consiste à reprendre* $\Xi_0(k)$ *et à superposer les deux processus de Poisson* $\mathcal{V}_k^C$ *et* $\mathcal{V}_k^T$.

$\Xi_{0,\rho}(k)$ *est le regroupement des deux processus de Poisson suivants :*
– *Pour* $x \in \{C, T\}$, $\mathcal{V}_k^x$ *le processus homogène de Poisson de taux* v_x.
– $\mathcal{V}_k^R := \mathcal{V}_k^A \cup \mathcal{V}_k^G$ *le processus homogène de Poisson de taux* $v_A + v_G$.

Ainsi, l'ensemble $\Xi_{0,\rho}(k)$ *consiste à reprendre* $\Xi_0(k)$ *et à superposer les deux processus de Poisson* $\mathcal{V}_k^A$ *et* $\mathcal{V}_k^G$.

Définition 6.2.8. *Pour un site* k*, à partir des marqueurs issus du processus* $\Xi_0(k)$ *et d'une condition initiale* $X_k(0)$*, on définit les marqueurs effectifs issus du processus* $\Xi_0(k)$ *le sous-ensemble des marqueurs de* $\Xi_0(k)$ *provoquant effectivement un mouvement depuis la condition initiale* $X_k(0)$.

Ce sous-ensemble ne dépend pas des marqueurs ou des conditions initiales présents sur les autres sites puisque ce sont les seuls marqueurs de transversion et que l'effectivité des mouvements associés à ces marqueurs ne dépend que du site k *considéré.*

On définit de même les marqueurs effectifs issus des processus $\Xi_{0,\pi}(k)$, $\Xi_{0,\eta}(k)$, $\Xi_{0,\rho}(k)$ *et des conditions initiales* $\pi_k(0)$, $\eta_k(0)$, $\rho_k(0)$.

Propriété 6.2.9. *Pour un site* k*, à partir de la connaissance de* $(X_k(t))_{t\in[0,T]}$ *(resp.* $(\pi_k(t))_{t\in[0,T]}$, $(\eta_k(t))_{t\in[0,T]}$, $(\rho_k(t))_{t\in[0,T]}$*), on déduit l'ensemble des marqueurs effectifs issus du processus* $\Xi_0(k)$ *(resp.* $\Xi_{0,\pi}(k)$, $\Xi_{0,\eta}(k)$, $\Xi_{0,\rho}(k)$*) en observant les instants de transversion.*

On établit ensuite la proposition suivante :

Proposition 6.2.10. *On décrit l'évolution* $(X_{1:m})$ *à partir de sa séquence initiale* $X_{1:m}(0)$ *et de la connaissance en chaque site* $k \in [\![1, m]\!]$ *des* 16 *processus de Poisson décrits dans la section 1.2.3 sur l'intervalle* $[0, T]$.

Pour décrire l'évolution $Z_{1:i}$ *à partir de la connaissance de ces processus et de cette séquence initiale, il suffit de connaître :*
– *La séquence initiale* $Z_{1:i}(0)$,
– *Les marqueurs associés aux processus* $\Xi_\rho(1)$ *et les marqueurs effectifs associés aux processus* $\Xi_{0,\rho}(1)$,
– *Les marqueurs associés aux processus* $\Xi_\rho(k)$ *et* $\Xi_\eta(k)$ *et* $\Xi_0(k)$ *pour* $k \in [\![2, i]\!]$,
– *Les marqueurs associés aux processus* $\Xi_\eta(i+1)$ *et les marqueurs effectifs associés aux processus* $\Xi_{0,\eta}(i + 1)$.

Démonstration. Supposons connaître la séquence initiale $Z_{1:i}(0)$, les marqueurs associés aux processus $\Xi_\rho(1)$ et les marqueurs effectifs associés aux processus $\Xi_{0,\rho}(1)$, les marqueurs

associés aux processus $\Xi_\rho(k)$ et $\Xi_\eta(k)$ et $\Xi_0(k)$ pour $k \in [\![2, i]\!]$ et les marqueurs associés aux processus $\Xi_\eta(i+1)$ et les marqueurs effectifs associés aux processus $\Xi_{0,\eta}(i+1)$.

On range par ordre croissant les marqueurs associés aux différents processus, associés aux instants notés $0 < t_1 < \ldots < t_r < T$. On note $t_{r+1} := T$ et on démontre par récurrence sur $l \in [\![1, r+1]\!]$ la propriété suivante :

$$H(l) : \text{L'évolution } Z_{1:i} \text{ est déterminée sur l'intervalle } [0, t_l[.$$

La propriété $H(1)$ est vérifiée car $Z_{1:i}(0)$ est connu et aucun marqueur n'est présent sur l'intervalle $[0, t_1[$.

Sous l'hypothèse $H(l)$, si le marqueur à l'instant t_l est sur l'un des sites k parmi $[\![2, i]\!]$ alors le mouvement associé est déterminé dans tous les cas puisque l'on connaît $\rho_{k-1}(t)$ et $\eta_{k+1}(t)$ (puisqu'aucun marqueur n'est présent à l'instant t aux sites $k-1$ et $k+1$).

Si le marqueur est sur le site 1, alors le mouvement est associé à un processus de $\Xi_\rho(1)$ ou de $\Xi_{0,\rho}(1)$ et ce mouvement ne fait intervenir sur les sites voisins uniquement $\eta_2(t)$ qui est déterminé. De même, si le marqueur est sur le site $i+1$, alors le mouvement est associé à un processus de $\Xi_\eta(i+1)$ ou de $\Xi_{0,\eta}(i+1)$ et ce mouvement ne fait intervenir sur les sites voisins uniquement $\rho_i(t)$.

Dans tous les cas, on en déduit l'évolution $Z_{1:i}$ sur l'intervalle $[0, t_l]$. Comme aucun marqueur n'est présent sur l'intervalle $]t_l, t_{l+1}[$, on en déduit l'évolution sur l'intervalle $[0, t_{l+1}[$, ce qui montre $H(l+1)$.

On a ainsi $H(r+1)$ vérifiée et l'évolution $Z_{1:i}$ est décrit sur $[0, T]$. □

6.2.4 Preuves des théorèmes 6.2.2 et 6.2.3

Preuve du théorème 6.2.2. Dans la preuve suivante, on démontre le théorème 6.2.2 à l'aide de construction de l'évolution encodée par processus de Poisson décrite dans la section 6.2.3.

Démonstration. On suppose connaître l'évolution $z_{1:i}$. D'après la proposition 6.2.10, cette évolution peut être vue comme la réalisation issue :

- d'une séquence initiale $Z_{1:i}(0)$,
- des marqueurs associés aux processus $\Xi_\rho(1)$ et des marqueurs effectifs associés à $\Xi_{0,\rho}(1)$,
- des marqueurs associés aux processus $\Xi_\rho(k)$ et $\Xi_\eta(k)$ et $\Xi_0(k)$ pour $k \in [\![2, i]\!]$,
- des marqueurs associés aux processus $\Xi_\eta(i+1)$ et des marqueurs effectifs associés à $\Xi_{0,\eta}(i+1)$.

Comme plusieurs jeux de marqueurs peuvent conduire à la même évolution $z_{1:i}$, ces marqueurs ne sont pas connus explicitement. Néanmoins, d'après la propriété 6.2.9, les marqueurs effectifs associés à $\Xi_{0,\eta}(i+1)$ sont connus explicitement.

On cherche maintenant à décrire l'évolution Z_{i+1} à partir de la connaissance de ces processus et de cette séquence initiale. Par la proposition 6.2.10, il suffit de connaître :

- Le couple encodé initial $Z_{i+1}(0)$,
- Les marqueurs associés aux processus $\Xi_\rho(i+1)$ et les marqueurs effectifs associés à $\Xi_{0,\rho}(i+1)$,
- Les marqueurs associés aux processus $\Xi_\eta(i+2)$ et les marqueurs effectifs associés à $\Xi_{0,\eta}(i+2)$.

La seule dépendance au passé est présente dans les marqueurs effectifs associés aux processus $\Xi_{0,\rho}(i+1)$. Comme marqueurs effectifs associés à $\Xi_{0,\eta}(i+1)$ sont connus explicitement, les instants des marqueurs effectifs associés aux processus $\Xi_{0,\rho}(i+1)$ correspondent aux instants des marqueurs effectifs de $\Xi_{0,\eta}(i+1)$. Par le principe de superposition des processus de Poisson, la valeur de ces marqueurs se déduit de la façon suivante :

- un marqueur A ou G sur $\Xi_{0,\eta}(i+1)$ correspond sur $\Xi_{0,\rho}(i+1)$ à un marqueur R,
- un marqueur Y sur $\Xi_{0,\eta}(i+1)$ correspond sur $\Xi_{0,\rho}(i+1)$ à un marqueur C avec probabilité $\frac{v_C}{v_C+v_T}$ et T sinon.

Ainsi, conditionnellement à $z_{1:i}$, l'évolution z_{i+1} se déduit de la connaissance de $z_{i+1}(0)$ et de l'évolution z_i (de laquelle on déduit par la propriété 6.2.9 les marqueurs effectifs associés au processus $\Xi_{0,\eta}(i+1)$).

Comme la loi à la racine est fixée, cela conclut la preuve.

□

Preuve du théorème 6.2.3. On veut maintenant montrer le caractère markovien explicite de l'évolution des couples Φ-encodés $(Z_i)_{i\in[\![1,m]\!]}$ donné par le théorème 6.2.3. On utilise les notations suivantes pour simplifier l'écriture des calculs, similaires à celles utilisées dans les notations 6.1.1 :

Notation 6.2.11. *Pour tout site i et tout instant t, on écrit :*

- *$z_i(t \to T)$ une évolution (ρ, η) aux sites $(i, i+1)$ de l'instant t à l'instant final T.*
- *$z_i(t)$ à la place de $Z_i(t) = z_i(t)$ dans l'écriture des probabilités.*
- *$z_i(t \to T)$ à la place de $(Z_i(s))_{s\in[t,T]} = z_i(t \to T)$.*

D'après le théorème 3.2.10, l'évolution de Z_i à partir d'un instant t ne dépend de la séquence non encodée $(X_0(t), \dots X_m(t))$ qu'à travers $(\rho(X_i)(t), \eta(X_{i+1})(t))$. On obtient en particulier le lemme suivant :

Lemme 6.2.12. *Pour $0 < s \leq t < T$, on a :*

$$P(z_{i-1}(t \to T)|z_{i-1}(t), z_i(s)) = P(z_{i-1}(t \to T)|z_{i-1}(t)).$$

Démonstration. Pour $s = t$, cela correspond à appliquer le dernier point du théorème 3.2.10. Sinon, on somme sur l'ensemble des couples $z_i(t)$ possibles :

$$\begin{aligned} P_0 := P(z_{i-1}(t \to T)|z_{i-1}(t), z_i(s)) &= \sum_{z_i(t)} P(z_{i-1}(t \to T), z_i(t)|z_{i-1}(t), z_i(s)) \\ &= \sum_{z_i(t)} P(z_{i-1}(t \to T)|z_{i-1}(t), z_i(s), z_i(t)) P(z_i(t)|z_{i-1}(t), z_i(s)). \end{aligned}$$

On utilise le caractère markovien en temps pour obtenir :

$$P_0 = \sum_{z_i(t)} P(z_{i-1}(t \to T)|z_{i-1}(t), z_i(t)) P(z_i(t)|z_{i-1}(t), z_i(s)).$$

On applique ensuite le dernier point du théorème 3.2.10 et on en déduit le résultat souhaité :

$$\begin{aligned} P_0 =& \sum_{z_i(t)} P(z_{i-1}(t \to T)|z_{i-1}(t)) P(z_i(t)|z_{i-1}(t), z_i(s)) \\ =& P(z_{i-1}(t \to T)|z_{i-1}(t)). \end{aligned}$$

□

On exprime ensuite dans la proposition suivante le taux de substitution instantanée de Z_i à un instant donné, conditionnellement au passé. On observe en particulier qu'il n'est pas nécessaire d'effectuer une intégration sur l'ensemble des sites futurs, contrairement à la structure markovienne de la section 6.1.5.

Proposition 6.2.13. *Pour toutes dates $0 < s < t < T$, pour tout site i et tous $z_{i-1}(s \to T)$, $z_i(s)$ et $z_i(t)$, on a :*

$$P(z_i(t)|z_{i-1}(0 \to T), z_i(s)) = P(z_i(t)|z_{i-1}(s \to t), z_i(s)).$$

Démonstration. Comme l'évolution est markovienne en temps, on a :

$$\begin{aligned} P_1 :=& P(z_i(t)|z_{i-1}(0 \to T), z_i(s)) = P(z_i(t)|z_{i-1}(s \to T), z_i(s)) \\ =& P(z_i(t)|z_{i-1}(s \to T), z_i(s)) \\ =& \frac{P(z_{i-1}(s \to T), z_i(s), z_i(t))}{P(z_{i-1}(s \to T), z_i(s))}. \end{aligned}$$

On écrit : $z_{i-1}(s \to T) = \{z_{i-1}(s \to t), z_{i-1}(t \to T)\}$, puis en utilisant le caractère markovien en temps :

$$\begin{aligned} P_1 =& \frac{P(z_{i-1}(t \to T)|z_{i-1}(s \to t), z_i(s), z_i(t))P(z_i(t)|z_{i-1}(s \to t), z_i(s))P(z_{i-1}(s \to t), z_i(s))}{P(z_{i-1}(t \to T)|z_{i-1}(s \to t), z_i(s))P(z_{i-1}(s \to t), z_i(s))} \\ =& \frac{P(z_{i-1}(t \to T)|z_{i-1}(t), z_i(t))}{P(z_{i-1}(t \to T)|z_{i-1}(t), z_i(s))} P(z_i(t)|z_{i-1}(s \to t), z_i(s)). \end{aligned}$$

On utilise enfin le lemme précédent aux valeurs s et t :

$$P_1 = \frac{P(z_{i-1}(t \to T)|z_{i-1}(t))}{P(z_{i-1}(t \to T)|z_{i-1}(t))} P(z_i(t)|z_{i-1}(s \to t), z_i(s)) = P(z_i(t)|z_{i-1}(s \to t), z_i(s)).$$

□

On écrit maintenant le raisonnement qui sera utilisé dans la preuve du théorème 6.2.3. On veut calculer le taux de substitution instantané de $z_i(t)$ à $z_i(t + \mathrm{d}t)$, connaissant $z_{i-1}(0 \to T)$. D'abord, on utilise la proposition 6.2.13 pour écrire :

$$P(z_i(t + \mathrm{d}t)|z_{i-1}(0 \to T), z_i(t)) = P(z_i(t + \mathrm{d}t)|z_{i-1}(t), z_{i-1}(t + \mathrm{d}t), z_i(t)).$$

Ensuite, on se sert du lemme 6.2.12 pour en déduire :

$$P(z_i(t + \mathrm{d}t)|z_{i-1}(0 \to T), z_i(t)) = \frac{P((z_{i-1}, z_i)(t + \mathrm{d}t)|(z_{i-1}, z_i)(t))}{P(z_{i-1}(t + \mathrm{d}t)|z_{i-1}(t))}.$$

La dernière écriture permet enfin d'appliquer le théorème 3.2.10 et d'expliciter les taux de sauts.

On utilise dans la preuve du théorème 6.2.3 la définition suivante, permettant d'indiquer les sauts possibles entre deux triplets encodés en terme de couples encodés :

Définition 6.2.14. *On rappelle (remarque 3.1.5) que :*

$$\mathcal{C} = \{RA, RG, RY, CA, CG, CY, TA, TG, TY\}.$$

On dit que $(z_1, z_1') \in \mathcal{C}^2$ *est compatible avec* $(z_2, z_2') \in \mathcal{C}^2$ *si* (z_1, z_2) *et* (z_1', z_2') *définissent bien des triplets encodés, et si* 0 *ou* 1 *lettre ne diffère entre ces deux triplets.*

Passons maintenant à la preuve du théorème 6.2.3.

Démonstration. On suppose que l'évolution $z_{i-1}(0 \to T)$ est connue pour un site i et on choisit un instant $t \in [0, T]$. Comme cette évolution a un nombre fini de sauts, si une substitution a lieu à l'instant t (respectivement si aucune substitution n'a lieu à l'instant t), alors dans un voisinage de t, l'évolution z_{i-1} a subi 1 saut (respectivement 0 saut). Soit $\varepsilon > 0$ assez petit pour que $z_{i-1}(t - \varepsilon/2 \to t + \varepsilon/2)$ ne comporte qu'une seule substitution (respectivement aucune substitution). On note par la suite $t- := t - \varepsilon/2$ et $t+ := t + \varepsilon/2$.

Soit $z_i(t-)$ et $z_i(t+)$ des couples Φ-encodés compatibles avec $z_{i-1}(t-)$ et $z_{i-1}(t+)$ au sens de la définition 6.2.14 (lorsque les éléments ne sont pas compatibles, le taux de substitution recherché est nul ou se comporte en $o(\varepsilon)$). On note alors pour toute date t_0 :

- $z_{i-1}(t_0) = (\rho_{i-1}(t_0), \eta_i(t_0))$ et $z_i(t_0) = (\rho_i(t_0), \eta_{i+1}(t_0))$,
- $x_i(t_0)$ l'unique nucléotide non encodé vérifiant : $\rho(x_i(t_0)) = \rho_i(t_0)$ et $\eta(x_i(t_0)) = \eta_i(t_0)$,
- $(z_{i-1}(t_0), z_i(t_0)) = (\rho_{i-1}(t_0), x_i(t_0), \eta_{i+1}(t_0))$ le triplet Φ-encodé associé aux dinucléotides encodés $z_{i-1}(t_0)$ et $z_i(t_0)$.

On applique la proposition 6.2.13 aux instants $t-$ et $t+$:

$$P(z_i(t+)|z_{i-1}(0 \to T), z_i(t-)) = P(z_i(t+)|z_{i-1}(t- \to t+), z_i(t-)).$$

Comme au plus une substitution a lieu au site $i-1$ sur l'intervalle $]t-, t+[$, on obtient :

$$P(z_i(t+)|z_{i-1}(0 \to T), z_i(t-)) = P(z_i(t+)|z_{i-1}(t-), z_{i-1}(t+), z_i(t-)) + o(\varepsilon).$$

On étudie en particulier le premier terme de la somme (on utilise le lemme 6.2.12 dans la deuxième égalité) :

$$P_2 := P(z_i(t+)|z_{i-1}(t+), z_{i-1}(t-), z_i(t-)) = \frac{P(z_{i-1}(t+), z_i(t+)|z_{i-1}(t-), z_i(t-))}{P(z_{i-1}(t+)|z_{i-1}(t-))}.$$

P_2 s'exprime comme le quotient du taux de substitution instantané en t de l'évolution encodée $(z_{i-1}, z_i) = (\rho_{i-1}, x_i, \eta_{i+1})$ divisé par le taux de substitution instantané en t de l'évolution encodée $z_{i-1} = (\rho_{i-1}, \eta_i)$. On rappelle (voir notations 6.0.5) que pour une matrice de taux de saut Q, on associe la matrice de transition instantanée $q^\varepsilon = I + \varepsilon Q$.

On utilise le théorème 3.2.10 pour exprimer le quotient P_2 :

$$\frac{q^\varepsilon_{,\eta_i(t-)}(\rho_{i-1}(t-), \rho_{i-1}(t+)) q^\varepsilon_{\rho_{i-1}(t-),\eta_{i+1}(t-)}(x_i(t-), x_i(t+)) q^\varepsilon_{\rho_i(t-),}(\eta_{i+1}(t-), \eta_{i+1}(t+))}{q^\varepsilon_{,\eta_i(t-)}(\rho_{i-1}(t-), \rho_{i-1}(t+)) q^\varepsilon_{\rho_{i-1}(t-),}(\eta_i(t-), \eta_i(t+))} + o(\varepsilon)$$

qui se réduit en :

$$\frac{q^\varepsilon_{\rho_{i-1}(t-),\eta_{i+1}(t-)}(x_i(t-), x_i(t+)) q^\varepsilon_{\rho_i(t-),}(\eta_{i+1}(t-), \eta_{i+1}(t+))}{q^\varepsilon_{\rho_{i-1}(t-),}(\eta_i(t-), \eta_i(t+))} + o(\varepsilon).$$

On distingue enfin cinq cas, suivant les valeurs fixées $\eta_i(t-)$ et $\eta_i(t+)$.

1. Si $\eta_i(t-) = \eta_i(t+) = Y$, on a $x_i(t-) \in \{C, T\}$ et $x_i(t+) \in \{C, T\}$, et on obtient le premier cas de la proposition.
2. Si $\eta_i(t-) = \eta_i(t+) \in \{A, G\}$, on a $x_i(t-) = x_i(t+) = R$ et on obtient le deuxième cas de la proposition.
3. Si $\eta_i(t-) = Y$ et $\eta_i(t+) \in \{A, G\}$, on obtient le troisième cas de la proposition.
4. Si $\eta_i(t-) \in \{A, G\}$ et $\eta_i(t+) = Y$, on obtient le quatrième cas de la proposition.
5. Enfin, si $\eta_i(t-) \in \{A, G\}$, $\eta_i(t+) \in \{A, G\}$ et $\eta_i(t+) \neq \eta_i(t-)$, l'évolution en z_i à cet instant est donnée par la matrice de substitution instantanée suivante :

$$\begin{array}{c} \\ RA \\ RG \\ RY \end{array} \begin{array}{c} \begin{array}{ccc} RA & RG & RY \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array}.$$

Par exemple, pour le cas 1 et avec $z_i(t-) = CA$ et $z_i(t+) = TA$, on obtient quelle que soit la valeur de $\rho_{i-1}(t-)$:

$$\begin{aligned} P_2 &= \frac{q^{\varepsilon}_{\rho_{i-1}(t-),A}(C,T) q^{\varepsilon}_{C,}(A,A)}{q^{\varepsilon}_{\rho_{i-1}(t-),}(Y,Y)} + o(\varepsilon) \\ &= \frac{(I + \varepsilon Q_{\rho_{i-1}(t-),A})(C,T) \times (I + \varepsilon Q_{C,})(A,A)}{(I + \varepsilon Q_{\rho_{i-1}(t-),})(Y,Y)} + o(\varepsilon) \\ &= \frac{\varepsilon Q_{\rho_{i-1}(t-),A}(C,T) \times (1 + \varepsilon Q_{C,}(A,A))}{(1 + \varepsilon Q_{\rho_{i-1}(t-),}(Y,Y))} + o(\varepsilon) \\ &= \varepsilon Q_{\rho_{i-1}(t-),A}(C,T) + o(\varepsilon) \\ &= \varepsilon(w_T + r_{CA \to TA}) + o(\varepsilon). \end{aligned}$$

□

6.2.5 Conditionnement par le dinucléotide encodé final

On étudie l'évolution précédente en rajoutant la connaissance du dinucléotide encodé final $Z_i(T)$. La forme obtenue est similaire à la matrice de taux de sauts de la proposition 6.1.15 et correspond une nouvelle fois à une h-transformée de Doob d'une chaîne de Markov.

La démonstration pour obtenir la matrice de taux de sauts associée est similaire et on obtient la proposition suivante :

Proposition 6.2.15. *On suppose que l'on connaît l'évolution sur les sites $Z_{1:i-1}$ et l'évolution de Z_i de l'instant initial jusqu'à l'instant t. On connaît aussi $Z_i(T) = z_i(T)$.*

Le taux de substitution de z vers z' du dinucléotide encodé Z_i à l'instant t est donné par :

$$\frac{G_{t \to T}(z', z_i(T))}{G_{t \to T}(z, z_i(T))} \hat{Q}_{t,\pi_i}(z, z').$$

6.2.6 Fonction de survie

Pour un instant t_0, on cherche à calculer la fonction de survie à partir de l'instant t_0 associée à l'évolution conditionnée décrite dans la proposition 6.2.15. Cette fonction de survie est définie pour tout $t > t_0$ et tout z dinucléotide encodé par la valeur $\bar{F}_{t_0}(t)(z)$.

On raisonne de la même façon que pour le calcul de la fonction de survie pour la structure de champ markovien (cf section 6.1.4) et on obtient la proposition suivante :

Proposition 6.2.16. *Pour $t \in [t_0, t_1[$ (où t_1 est défini selon la définition 6.2.4) et z dinucléotide encodé, la valeur de la fonction de survie $\bar{F}_{t_0}(t)(z)$ est donnée par :*

$$\exp\left(-(t-t_0)\sum_{z'\neq z}\hat{Q}_{t_0,\pi_i}(z,z')\right)\frac{G(t)(z,z_T)}{G(t_0)(z,z_T)}.$$

6.3 Structure basée sur le π-encodage

On a vu dans la section 6.2 une structure markovienne du modèle RN95+YpR basée sur la structure en termes de dinucléotides encodés (ρ, η). On décrit ici une structure décrivant les évolutions encodées sous la forme (ρ, η) mais cette fois conditionnellement aux évolutions π-encodées (voir définition 3.2.1). On obtient une structure de suite indépendante conditionnellement à ces évolutions π-encodées, que l'on explicite.

Pour un site j et un instant t, on note $\pi_j(t-) = \lim_{\varepsilon\to 0}\pi_j(t-\varepsilon)$.

On a la structure d'indépendance suivante :

Proposition 6.3.1. *On suppose connue l'évolution π-encodée pour les sites i et $i+1$ et on fixe une séquence initiale $(Z_k(0))_{k\in[\![1,m-1]\!]}$. Alors conditionnellement à $(Z_k(0))_k$, à π_i et à π_{i+1}, la variable Z_i est indépendante des variables $(Z_k)_{k\in[\![1,i-1]\!]}$ et $(Z_k)_{k\in[\![i+1,m-1]\!]}$.*

On suppose de plus que l'évolution de Z_i est connue de l'instant initial jusqu'à l'instant t. Conditionnellement à ces évolutions, la dépendance dans l'écriture de l'évolution de Z_i à l'instant t se fait seulement à travers $\pi_i(t-), \pi_i(t), \pi_{i+1}(t-)$ et $\pi_{i+1}(t)$. On décrit l'évolution à l'aide de matrices de taux de sauts et de matrices de substitution instantanées.

Lorsque $\pi_i(t-) = \pi_i(t)$ et $\pi_{i+1}(t-) = \pi_{i+1}(t)$, l'évolution s'exprime à l'aide de matrices de taux de sauts suivantes (chaque terme non indiqué est tel que la somme de chaque ligne soit égale à zéro) :

- *Si $\pi_i(t) = R$ et $\pi_{i+1}(t) = Y$, alors $(\rho_i(t), \eta_{i+1}(t)) = RY$ et l'évolution est décrite par :*

$$\begin{array}{c} \\ RY \end{array}\begin{array}{c} \begin{array}{c} RY \end{array} \\ \begin{pmatrix} 0 \end{pmatrix}. \end{array}$$

- *Si $\pi_i(t) = R$ et $\pi_{i+1}(t) = R$, alors $(\rho_i(t), \eta_{i+1}(t)) \in \{RA, RG\}$ et l'évolution est décrite par :*

$$\begin{array}{c} \\ RA \\ RG \end{array}\begin{array}{c} \begin{array}{cc} RA & RG \end{array} \\ \begin{pmatrix} . & w_G \\ w_A & . \end{pmatrix}. \end{array}$$

- *Si $\pi_i(t) = Y$ et $\pi_{i+1}(t) = Y$, alors $(\rho_i(t), \eta_{i+1}(t)) \in \{CY, TY\}$ et l'évolution est décrite par :*

$$\begin{array}{c} \\ CY \\ TY \end{array}\begin{array}{c} \begin{array}{cc} CY & TY \end{array} \\ \begin{pmatrix} . & w_T \\ w_C & . \end{pmatrix}. \end{array}$$

– *Si* $\pi_i(t) = Y$ *et* $\pi_{i+1}(t) = R$, *alors* $(\rho_i(t), \eta_{i+1}(t)) \in \{CA, CG, TA, TG\}$ *et l'évolution est décrite par :*

$$\begin{array}{c} \\ CA \\ CG \\ TA \\ TG \end{array} \begin{array}{c} \begin{array}{cccc} CA & CG & TA & TG \end{array} \\ \begin{pmatrix} . & w_G + r_{CA\to CG} & w_T + r_{CA\to TA} & 0 \\ w_A + r_{CG\to CA} & . & 0 & w_T + r_{CG\to TG} \\ w_C + r_{TA\to CA} & 0 & . & w_G + r_{TA\to TG} \\ 0 & w_C + r_{TG\to CG} & w_A + r_{TG\to TA} & . \end{pmatrix} \end{array}.$$

Lorsque $\pi_i(t-) \neq \pi_i(t)$ *ou* $\pi_{i+1}(t-) \neq \pi_{i+1}(t)$, *l'évolution s'exprime à l'aide de matrices de substitution instantanées suivantes :*

– *Si* $\pi_i(t-) = R \neq Y = \pi_i(t)$ *et* $\pi_{i+1}(t-) = \pi_{i+1}(t) = Y$, *la matrice est donnée par :*

$$\begin{array}{c} \\ RY \end{array} \begin{array}{c} \begin{array}{cc} CY & TY \end{array} \\ \begin{pmatrix} v_C & v_T \end{pmatrix} \end{array} \times \frac{1}{v_C + v_T}.$$

– *Si* $\pi_i(t-) = \pi_i(t) = R$ *et* $\pi_{i+1}(t-) = Y \neq R = \pi_{i+1}(t)$, *la matrice est donnée par :*

$$\begin{array}{c} \\ RY \end{array} \begin{array}{c} \begin{array}{cc} RA & RG \end{array} \\ \begin{pmatrix} v_A & v_G \end{pmatrix} \end{array} \times \frac{1}{v_A + v_G}.$$

– *Si* $\pi_i(t-) = R \neq Y = \pi_i(t)$ *et* $\pi_{i+1}(t-) = \pi_{i+1}(t) = R$, *la matrice est donnée par :*

$$\begin{array}{c} \\ RA \\ RG \end{array} \begin{array}{c} \begin{array}{cccc} CA & CG & TA & TG \end{array} \\ \begin{pmatrix} v_C & 0 & v_T & 0 \\ 0 & v_C & 0 & v_T \end{pmatrix} \end{array} \times \frac{1}{v_C + v_T}.$$

– *Si* $\pi_i(t-) = \pi_i(t) = Y$ *et* $\pi_{i+1}(t-) = Y \neq R = \pi_{i+1}(t)$, *la matrice est donnée par :*

$$\begin{array}{c} \\ CY \\ TY \end{array} \begin{array}{c} \begin{array}{cccc} CA & CG & TA & TG \end{array} \\ \begin{pmatrix} v_A & v_G & 0 & 0 \\ 0 & 0 & v_A & v_G \end{pmatrix} \end{array} \times \frac{1}{v_A + v_G}.$$

– *Si* $\pi_i(t-) = \pi_i(t) = R$ *et* $\pi_{i+1}(t-) = R \neq Y = \pi_{i+1}(t)$, *la matrice est donnée par :*

$$\begin{array}{c} \\ RA \\ RG \end{array} \begin{array}{c} \begin{array}{c} RY \end{array} \\ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \end{array}.$$

– *Si* $\pi_i(t-) = Y \neq R = \pi_i(t)$ *et* $\pi_{i+1}(t-) = \pi_{i+1}(t) = Y$, *la matrice est donnée par :*

$$\begin{array}{c} \\ CY \\ TY \end{array} \begin{array}{c} \begin{array}{c} RY \end{array} \\ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \end{array}.$$

– *Si* $\pi_i(t-) = Y \neq R = \pi_i(t)$ *et* $\pi_{i+1}(t-) = \pi_{i+1}(t) = R$, *la matrice est donnée par :*

$$\begin{array}{c} \\ CA \\ CG \\ TA \\ TG \end{array} \begin{array}{c} \begin{array}{cc} RA & RG \end{array} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array}.$$

– *Si* $\pi_i(t-) = \pi_i(t) = Y$ *et* $\pi_{i+1}(t-) = R \neq Y = \pi_{i+1}(t)$*, la matrice est donnée par :*

$$\begin{array}{c} \\ CA \\ CG \\ TA \\ TG \end{array} \begin{array}{c} \begin{array}{cc} CY & TY \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \end{array}.$$

Démonstration. On raisonne de manière analogue aux preuves des théorèmes 6.2.2 et 6.2.3. On suppose connaître l'évolution π-encodée pour les sites $i-1$ et i. En reprenant les notations utilisées pour la construction de l'évolution encodée par processus de Poisson (section 6.2.3), on en déduit les marqueurs effectifs issus des processus $\Xi_{0,\pi}(i)$ et $\Xi_{0,\pi}(i+1)$. Par désambigüisation, on construit les marqueurs effectifs issus des processus $\Xi_{0,\rho}(i)$ et $\Xi_{0,\eta}(i+1)$. Indépendamment, on pose les marqueurs issus des processus $\Xi_\rho(i)$ et $\Xi_\eta(i+1)$. À partir de la condition initiale, on en déduit par la proposition 6.2.10 que cela suffit pour décrire l'évolution (Z_i) et on en déduit la première partie de la proposition 6.3.1.

Pour l'écriture explicite des transitions, on raisonne comme dans la preuve du théorème 6.2.3. En reprenant les notations de cette preuve, on écrit la transition instantanée comme :

$$P(z_i(t+)|z_i(t-), \pi_i(t+), \pi_{i+1}(t+)) = \frac{P(z_i(t+)|z_i(t-))}{P(\pi_i(t+), \pi_{i+1}(t+)|z_i(t-))}.$$

Par les propriétés d'indépendance des évolutions (π_i), on écrit ensuite :

$$P(z_i(t+)|z_i(t-), \pi_i(t+), \pi_{i+1}(t+)) = \frac{P(z_i(t+)|z_i(t-))}{P(\pi_i(t+)|\pi_i(t-))P(\pi_{i+1}(t+)|\pi_{i+1}(t-))}.$$

Ensuite, en notant $q_\pi^\varepsilon = I + \varepsilon Q_\pi$ la matrice de transition instantanée associée à la matrice de taux de sauts Q_π (voir définition 3.2.4), on réécrit la transition comme :

$$\frac{q^\varepsilon_{,\eta_{i+1}(t-)}(\rho_i(t-), \rho_i(t+)) q^\varepsilon_{\rho_i(t-),}(\eta_{i+1}(t-), \eta_{i+1}(t+))}{q_\pi^\varepsilon(\pi_i(t-), \pi_i(t+)) q_\pi^\varepsilon(\pi_i(t-), \pi_i(t+))} + o(\varepsilon).$$

On distingue enfin les différents cas, suivant les valeurs fixées $\pi_i(t-)$, $\pi_i(t+)$, $\pi_{i+1}(t-)$ et $\pi_{i+1}(t+)$, de façon analogue à la preuve du théorème 6.2.3. □

Remarque 6.3.2.

– *On obtient des expressions de la fonction de survie similaires à celles obtenues dans la section 6.2.*
– *La proposition 6.3.1 suppose la connaissance de l'évolution π-encodée pour chaque site, qui n'est pas connu directement en général. Néanmoins, lorsque l'évolution n'est pas conditionnée par une séquence finale, l'évolution π-encodée est explicite puisque chaque site de l'évolution suit de manière indépendante la matrice de taux de sauts Q_π. En particulier, cela fournit une méthode pour simuler de façon exacte la loi stationnaire, décrite dans la section 8.3.*

6.4 Description des structures sur le modèle complet

Dans les structures décrites aux sections 6.1, 6.2 et 6.3, nous avons considéré l'évolution issue d'un modèle d'évolution M inclus dans la classe RN95+YpR, à racine fixée et sur l'arbre réduit à une arête (évolution de séquence à séquence).

On va voir que l'on obtient des structures similaires pour les modèles complets $(\mathsf{R}, \mathsf{T}, \mathsf{M})$ décrits dans la section 1.3.2 à l'aide d'une loi à la racine R, d'un arbre enraciné T et d'un modèle d'évolution M. Ces généralisations sont décrites dans les deux sections suivantes 6.4.1 et 6.4.2, où on rappelle en outre l'algorithme par récurrence de Felsenstein [42] permettant de déduire efficacement la vraisemblance de séquences associées aux feuilles d'un arbre à partir de la connaissance des vraisemblances de séquence à séquence.

6.4.1 Description de l'évolution lorsque la racine n'est pas constante

On considère tout d'abord le cas des structures basées sur les couples Φ-encodés (sections 6.2 et 6.3). On suppose que la loi à la racine vérifie la propriété suivante :

Hypothèse 6.4.1. *La loi de la racine $(Z_i(0))_{i\in[\![1,m]\!]}$ vérifie pour tout $i \in [\![1, m]\!]$ et conditionnellement aux évolutions $Z_{1:i}$,*

$$P(Z_{i+1}(0)|Z_{1:i}) = P(Z_{i+1}(0)|Z_{1:i}(0)).$$

Remarque 6.4.2. *On observe que l'on peut choisir pour la loi à la racine la loi stationnaire du modèle M considéré pour l'évolution, ou une approximation markovienne (avec un nombre de pas quelconque) de cette loi stationnaire.*

On a la proposition suivante :

Proposition 6.4.3. *On suppose que l'évolution suit un modèle M et une loi à la racine régie par R.*

Sous l'hypothèse 6.4.1, pour tout site $i \in [\![1, m]\!]$, l'évolution du couple Φ-encodée Z_{i+1} conditionnellement aux évolutions $Z_{1:i}$ vérifie la propriété suivante :

$$P(Z_{i+1}|Z_{1:i}) = \sum_{Z_{i+1}(0)} P(Z_{i+1}|Z_i, Z_{i+1}(0))P(Z_{i+1}(0)|Z_{1:i}(0)).$$

Cette proposition permet de séparer le calcul de la racine et celle de l'évolution conditionnée, en considérant tout d'abord la valeur de la racine $Z_{i+1}(0)$ puis celle à racine fixée de l'évolution Z_{i+1} conditionnellement à Z_i.

Démonstration. On écrit directement avec l'hypothèse 6.4.1 et le théorème 6.2.2 :

$$\begin{aligned} P(Z_{i+1}|Z_{1:i}) &= \sum_{Z_{i+1}(0)} P(Z_{i+1}|Z_{1:i}, Z_{i+1}(0))P(Z_{i+1}(0)|Z_{1:i}) \\ &= \sum_{Z_{i+1}(0)} P(Z_{i+1}|Z_i, Z_{i+1}(0))P(Z_{i+1}(0)|Z_{1:i}(0)). \end{aligned}$$

□

Dans le cas où on a de plus conditionné par le couple encodé final $Z_{i+1}(T)$, on a la proposition similaire suivante qui permet également de découpler le calcul de l'évolution au site 0 et le calcul de l'évolution à racine fixée.

Proposition 6.4.4. *Sous l'hypothèse 6.4.1, pour tout site $i \in [\![1, m]\!]$, l'évolution du couple Φ-encodée Z_{i+1} conditionnellement aux évolutions $Z_{1:i}$ et au site final $Z_{i+1}(T)$ vérifie la propriété suivante :*

$$P(Z_{i+1}|Z_{1:i}, Z_{i+1}(T)) = \sum_{Z_{i+1}(0)} P(Z_{i+1}|Z_i, Z_{i+1}(0), Z_{i+1}(T)) \frac{P(Z_{i+1}(T)|Z_{i+1}(0), Z_i)P(Z_{i+1}(0)|Z_{1:i}(0))}{\sum_{\tilde{Z}_{i+1}(0)} P(Z_{i+1}(T)|\tilde{Z}_{i+1}(0), Z_i)P(\tilde{Z}_{i+1}(0)|Z_{1:i}(0))}.$$

Démonstration. On reprend la preuve de la proposition 6.4.3 et on calcule en particulier le terme $P(Z_{i+1}(0)|Z_{1:i}, Z_{i+1}(T))$ de la façon suivante :

$$P(Z_{i+1}(0)|Z_{1:i}, Z_{i+1}(T)) = \frac{P(Z_{i+1}(T)|Z_{i+1}(0), Z_i)P(Z_{i+1}(0)|Z_{1:i}(0))}{\sum_{\tilde{Z}_{i+1}(0)} P(Z_{i+1}(T)|\tilde{Z}_{i+1}(0), Z_i)P(\tilde{Z}_{i+1}(0)|Z_{1:i}(0))}.$$

□

On considère maintenant le cas de la structure de champ markovien (section 6.1). On obtient de façon analogue l'hypothèse et la proposition suivantes :

Hypothèse 6.4.5. *La loi de la racine $(X_i(0))_{i\in[\![1,m]\!]}$ vérifie pour tout $i \in [\![1, m]\!]$ et conditionnellement aux évolutions $X_{1:i-1}, X_{i+1:m}$,*

$$P(X_i(0)|X_{1:i-1}, X_{i+1:m}) = P(X_i(0)|X_{1:i-1}(0), X_{i+1:m}(0)).$$

Proposition 6.4.6. *On suppose que l'évolution suit un modèle* M *et une loi à la racine régie par* R*.*

Sous l'hypothèse 6.4.5, pour tout site $i \in [\![1, m]\!]$, on a :

$$P(X_i|X_{1:i-1}, X_{i+1:m}) = \sum_{X_i(0)} P(X_i|X_i(0), X_{i-1}, X_{i+1})P(X_i(0)|X_{1:i-1}(0), X_{i+1:m}(0)),$$

$$P(X_i|X_{1:i-1}, X_{i+1:m}, X_i(T)) = \sum_{X_i(0)} P(X_i|X_i(0), X_i(T), X_{i-1}, X_{i+1}) \times \frac{P(X_i(T)|X_{i-1}, X_{i+1}, X_i(0))P(X_i(0)|X_{1:i-1}(0), X_{i+1:m}(0))}{\sum_{\tilde{X}_i(0)} P(X_i(T)|X_{i-1}, X_{i+1}, \tilde{X}_i(0))P(\tilde{X}_i(0)|X_{1:i-1}(0), X_{i+1:m}(0))}.$$

6.4.2 Description de l'évolution sur un arbre

On reprend la définition 1.3.7 du modèle complet : on considère une évolution continue sur un espace d'états fini S ainsi qu'un arbre enraciné et les différentes longueurs des branches. On suppose connaître pour chaque arête (associée à son nœud fils v), pour tous $x, x' \in S$ et tous t_0, t instants sur l'arête :

- $Q_{t,v}(x, x')$ la probabilité instantanée de substitution en t de $x \in S$ vers x'.
- $G_{t\to v}(x, x')$ la matrice de transition de x en t vers x' à l'instant correspondant au nœud v.

Remarque 6.4.7. *L'évolution se met sous cette forme pour les trois structures précédemment décrites.*

Pour (v, v') une arête orientée, on note $G_{v\to v'}$ la matrice de transition de l'instant correspondant au nœud v à l'instant correspondant au nœud v'.

Pour tout nœud v, on note par $l(v)$ l'ensemble des nœuds feuilles de l'arbre issus de v. En utilisant la propriété d'indépendance le long des arêtes de l'arbre (voir la construction de l'évolution dans la définition 1.3.7), on obtient la propriété suivante :

Propriété 6.4.8. *On obtient par récurrence la matrice de transition sur l'arbre de la façon suivante, pour chaque nœud v, $x \in S$ et $y \in S^{\sharp l(v)}$:*

- *Si v est une feuille, $G_{v\to l(v)}(x, y) = \mathbf{1}(x = y)$.*
- *Sinon, on note v_1 et v_2 les deux nœuds fils de v et on pose :*

$$G_{v\to l(v)}(x,y) = \left(\sum_{x_1\in S} G_{v\to v_1}(x,x_1)G_{v_1\to l(v_1)}(x_1,y)\right)\left(\sum_{x_2\in S} G_{v\to v_2}(x,x_2)G_{v_2\to l(v_2)}(x_2,y)\right).$$

Enfin, on définit pour toute arête de nœud fils v, pour tout instant t sur cette arête, et tous $x \in S$, $y \in S^{\sharp l(v)}$:

$$G_{t,v\to l(v)}(x,y) = \sum_{x_1\in S} G_{t\to v}(x,x_1)G_{v\to l(v)}(x_1,y).$$

Cette dernière matrice est la matrice de transition de x en t vers y aux instants feuilles.

La taux de substitution conditionné est alors décrit par la formule suivante :

Propriété 6.4.9. *Sachant les feuilles $x(T)$, le taux de substitution de x vers y à l'instant t est donnée par :*

$$\frac{G_{t,v\to l(v)}(y,x(T))}{G_{t,v\to l(v)}(x,x(T))}Q_{t,v}(x,y).$$

Pour le calcul effectif de la vraisemblance de séquences observées associées aux feuilles de l'arbre, on utilise l'algorithme de Felsenstein proposé dans [42] (section *Computing the Likelihood of a Tree*). On choisit alors à chaque étape un nœud v vérifiant pour tout nœud v' issu de ce nœud que $G_{v'\to l(v')}(., y)$ est connu. On peut à chaque étape choisir un tel v, et le calcul de $G_{v\to l(v)}(., y)$ s'effectue grâce à l'égalité énoncée dans la propriété 6.4.8. Un exemple sur un arbre à 5 feuilles est proposé dans [42].

Chapitre 7

Propriétés du maximum de vraisemblance des observations

On souhaite montrer dans ce chapitre la consistance et la normalité asymptotique de l'estimateur du maximum de vraisemblance dans le cas d'observations issues d'un modèle RN95+YpR.

On se limite au cas où la topologie de l'arbre et les différentes longueurs de branches T_0 sont fixées, ainsi que la loi à la racine R_0. On considère un ensemble de modèles Θ inclus dans la classe de modèles globaux RN95+YpR. On choisit $\theta_0 \in \Theta$ et on se donne pour tout m un jeu de séquences observées de longueurs m issu de ce modèle. Par abus de notation et pour simplifier l'écriture, on note $\Phi_m(x(T))$ l'ensemble des séquences Φ-encodées associées aux feuilles de l'arbre (l désigne dans l'égalité suivante le nombre de feuilles de l'arbre) :

$$\Phi_m(x(T)) = (\rho(x_1)(T), x_2(T), \ldots, x_{m-1}(T), \eta(x_m)(T)) \in \left(\{C,T,R\} \times \mathcal{A}^{m-2} \times \{A,G,Y\}\right)^l.$$

On cherche à estimer θ_0 par la méthode du maximum de vraisemblance. En désignant par $L_\theta(\Phi_m(x(T)))$ la vraisemblance des observations pour un modèle $\theta \in \Theta$, cela correspond à étudier un estimateur du maximum de vraisemblance :

$$\hat{\theta}_m := \mathrm{argmax}_{\theta\in\Theta}\, L_\theta(\Phi_m(x(T))).$$

On veut montrer sous certaines conditions sur l'ensemble Θ que cet estimateur est :
- consistant, c'est-à-dire que $\hat{\theta}_m \to_{m\to+\infty} \theta_0$ P_{θ_0}-p.s.
- asymptotiquement normal, c'est-à-dire qu'il existe une matrice symétrique définie positive $\mathcal{I}(\theta_0)^{-1}$ telle que l'on ait en loi :

$$m^{1/2}(\hat{\theta}_m - \theta_0) \to_{m\to+\infty} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1}).$$

Pour cela, on utilise la structure markovienne décrite dans la section 6.2 et rappelée maintenant. On considère un modèle global dont le modèle d'évolution appartient à RN95+YpR et on suppose que l'évolution est effectuée sur m sites Φ-encodés. L'évolution est régie par les variables :

$$(\rho(X_1), X_2, \ldots, X_{m-1}, \eta(X_m)).$$

On pose alors pour $i \in [\![1, m-1]\!]$: $Z_i = (\rho(X_i), \eta(X_{i+1}))$. D'après le théorème 6.2.2, l'évolution spatiale des Z_i est markovienne.

Cette propriété markovienne induit une structure de chaîne de Markov cachée (voir définition dans la section 7.1.1), avec les variables cachées $(Z_i)_{i\in[\![1,m-1]\!]}$ et les variables observées $(Z_i(T))_{i\in[\![1,m-1]\!]}$. Notons que conditionnellement aux variables cachées, les variables observées sont ici déterministes.

$$\begin{array}{ccccccccc} Z_1 & \longrightarrow & Z_2 & \longrightarrow & Z_3 & \longrightarrow & \dots & \longrightarrow & Z_{m-1} \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\ Z_1(T) & & Z_2(T) & & Z_3(T) & & \dots & & Z_{m-1}(T) \end{array} \tag{7.1}$$

On cherche à exploiter cette structure de chaîne de Markov cachée à travers des théorèmes de consistance et de normalité asymptotique du maximum de vraisemblance spécifiques aux chaînes de Markov cachées énoncés dans la section 7.1. En particulier, on énonce le théorème 7.1.9 issu de [21] (théorème 12.5.7 dans le livre).

Pour se placer dans la théorie existante des chaînes de Markov cachées, on a besoin de formuler différemment la structure markovienne des $(Z_i)_{i\in[\![1,m-1]\!]}$, et en particulier d'établir son noyau de transition par rapport à une mesure de référence. On décrit cette reformulation dans la section 7.2 de séquence à séquence puis sur un arbre.

Enfin, on cherche à appliquer dans la section 7.3 les théorèmes généraux de la section 7.1 à notre chaîne de Markov cachée particulière reformulée dans la section 7.2.

Pour appliquer le théorème 7.1.9, on impose des hypothèses génériques (hypothèses 7.3.1 et 7.3.2) de compacité et d'identifiabilité de la classe Θ de modèles considérés. Néanmoins, ces hypothèses ne permettent pas d'appliquer directement le théorème 7.1.9 puisque des difficultés apparaissent liées au fait que toutes les transitions entre Z_i et Z_{i+1} ne sont pas possibles, et à l'utilisation de variables observées déterministes conditionnellement aux variables cachées (voir (7.1)).

On effectue alors dans la suite de la section 7.3 un travail d'adaptation permettant de conclure finalement, sous les hypothèses de compacité et d'identifiabilité de Θ, que l'estimateur du maximum de vraisemblance vérifie les propriétés de consistance et de normalité asymptotique souhaitées.

7.1 Théorèmes limites pour les chaînes de Markov cachées

On définit ce qu'est une chaîne de Markov cachée avant d'énoncer un théorème générique issu de [21] de consistance et de normalité asymptotique du maximum de vraisemblance pour les chaînes de Markov cachées.

7.1.1 Définition d'une chaîne de Markov cachée et propriétés des chaînes de Markov

Définition d'une chaîne de Markov cachée.

On veut définir une chaîne de Markov cachée, c'est-à-dire de manière générale une suite de variables aléatoires $(\underset{\star}{X}_i, \underset{\star}{Y}_i)_{i\in 1:m}$ vérifiant les propriétés suivantes :

- la suite $(\underset{\star}{X}_i)_{i\in 1:m}$ est une chaîne de Markov,
- les variables $(\underset{\star}{Y}_i)_{i\in 1:m}$ sont indépendantes conditionnellement à $(\underset{\star}{X}_i)_{i\in[\![1,m]\!]}$ et

– pour tout i, $\underset{\star}{Y}_i$ ne dépend que de $\underset{\star}{X}_i$ conditionnellement à $(\underset{\star}{X}_i)_{i\in[\![1,m]\!]}$.

Formellement, on définit une chaîne de Markov cachée de la façon suivante (cette définition est issue de [21]) :

Définition 7.1.1. *Soit $(\underset{\star}{\mathsf{X}}, \underset{\star}{\mathcal{X}})$ et $(\underset{\star}{\mathsf{Y}}, \underset{\star}{\mathcal{Y}})$ deux espaces mesurables, $\underset{\star}{Q}$ un noyau de transition de $(\underset{\star}{\mathsf{X}}, \underset{\star}{\mathcal{X}})$ et $\underset{\star}{G}$ un noyau de transition de $(\underset{\star}{\mathsf{X}}, \underset{\star}{\mathcal{X}})$ vers $(\underset{\star}{\mathsf{Y}}, \underset{\star}{\mathcal{Y}})$. On considère le noyau de transition sur $(\underset{\star}{\mathsf{X}} \times \underset{\star}{\mathsf{Y}}, \underset{\star}{\mathcal{X}} \otimes \underset{\star}{\mathcal{Y}})$ défini pour tous $(x, y) \in \underset{\star}{\mathsf{X}} \times \underset{\star}{\mathsf{Y}}$ et $C \in \underset{\star}{\mathcal{X}} \otimes \underset{\star}{\mathcal{Y}}$ par :*

$$T\left((x,y),C\right) = \int\int_C \underset{\star}{Q}(x,dx')\underset{\star}{G}(x',dy').$$

La chaîne de Markov $(\underset{\star}{X}_i, \underset{\star}{Y}_i)_{i\in\mathbb{N}^*}$ de noyau de transition T et de loi initiale $\nu \otimes \underset{\star}{G}$, où ν est une mesure de probabilité sur $(\underset{\star}{\mathsf{X}}, \underset{\star}{\mathcal{X}})$ est appelée chaîne de Markov cachée.

La chaîne des m premiers pas $(\underset{\star}{X}_i, \underset{\star}{Y}_i)_{i\in[\![1,m]\!]}$ est appelée chaîne de Markov cachée sur $[\![1, m]\!]$.

Propriétés des chaînes de Markov.

On définit ici les notions d'irréductibilité et de positivité pour les chaînes de Markov qui seront utilisées dans la section 7.1.2. Ces définitions proviennent de [21], sections 14.2.1 et 14.2.3.

Définition 7.1.2. *Irréductibilité. Le noyau de transition $\underset{\star}{Q}$ associé à une chaîne de Markov $(\underset{\star}{X}_i)_{i\in\mathbb{N}^*}$ est dit irréductible s'il existe une mesure ψ sur $(\underset{\star}{\mathsf{X}}, \underset{\star}{\mathcal{X}})$ telle que pour tout $A \in \underset{\star}{\mathcal{X}}$ vérifiant $\psi(A) > 0$, pour tout $x \in \underset{\star}{\mathsf{X}}$, la probabilité partant de x que le premier temps de retour dans A soit fini est strictement positive.*

Définition 7.1.3. *Positivité. Un noyau de transition $\underset{\star}{Q}$ irréductible est dit positif s'il admet une mesure* de probabilité *invariante.*

7.1.2 Théorèmes de consistance et de normalité asymptotique

Différents théorèmes de consistance et de normalité asymptotique du maximum de vraisemblance pour les chaînes de Markov cachées ont été développés. Une brève bibliographie est disponible dans la section 12.7 de [21]. Parmi ces théorèmes, on ne peut utiliser que ceux utilisant un espace caché mesurable général – qui n'est pas forcément supposé compact. Deux théorèmes sont alors utilisables :

1. le théorème de consistance issu de [34],
2. le théorème de consistance et de normalité asymptotique de [21].

Le théorème issu de [34] peut être effectivement utilisé pour montrer la consistance de l'estimateur. Elle considère directement que la loi initiale (de la première variable cachée $X_1 = Z_1$) est la loi stationnaire en espace le long de la séquence, ce qui est le cas pour notre modèle. En effet, si on considère la structure de la section 6.3 basée sur le π-encodage, sans conditionnement par des observations, les variables $(\pi(X_i))_i$ sont indépendantes, puis conditionnellement à ces variables, les variables $(Z_i)_i$ sont indépendantes). Cela permet d'éviter d'être confronté au problème d'oubli de la condition initiale et évite donc d'avoir à adapter la preuve du théorème.

Néanmoins, on choisit d'utiliser le théorème issu [21] énoncé ici comme le théorème 7.1.9. Celui-ci permet d'obtenir à la consistance mais également la normalité asymptotique de

l'estimateur du maximum de vraisemblance. On rappelle que la chaîne de Markov cachée considérée est paramétrée par $\theta_0 \in \Theta$, où Θ est un sous-ensemble compact de $\mathbb{R}^d$ (pour un certain $d \geq 1$). On considère les hypothèses suivantes :

Hypothèse 7.1.4.

- *Il existe une mesure de probabilité $\underset{\star}{\lambda}$ sur $(\underset{\star}{\mathsf{X}}, \underset{\star}{\mathcal{X}})$ telle que pour tout $x \in \underset{\star}{\mathsf{X}}$ et tout $\theta \in \Theta$, $\underset{\star}{Q}_\theta(x,.) \ll \underset{\star}{\lambda}$ avec une densité de transition $\underset{\star}{q}_\theta$. Cela correspond à écrire $\underset{\star}{Q}_\theta(x,A) = \int \underset{\star}{q}_\theta(x,x')\underset{\star}{\lambda}(dx')$ pour $A \in \underset{\star}{\mathcal{X}}$, avec $\underset{\star}{q}_\theta$ à valeurs dans $\mathbb{R}^+$ et $\underset{\star}{\mathcal{X}} \otimes \underset{\star}{\mathcal{Y}}$-mesurable.*
- *Il existe une mesure de probabilité $\underset{\star}{\mu}$ sur $(\underset{\star}{\mathsf{Y}}, \underset{\star}{\mathcal{Y}})$ telle que pour tout $x \in \underset{\star}{\mathsf{X}}$ et tout $\theta \in \Theta$, $\underset{\star}{G}_\theta(x,.) \ll \underset{\star}{\mu}$ avec une densité de transition $\underset{\star}{g}_\theta$. On écrit alors, $\underset{\star}{G}_\theta(x,A) = \int \underset{\star}{g}_\theta(x,y)\underset{\star}{\mu}(dy)$ pour $A \in \underset{\star}{\mathcal{Y}}$.*
- *Pour tout $\theta \in \Theta$, $\underset{\star}{Q}_\theta$ est positif, c'est-à-dire $\underset{\star}{Q}_\theta$ est irréductible et admet une loi invariante (nécessairement unique) notée π_θ.*

Hypothèse 7.1.5.

- *La densité de transition $\underset{\star}{q}_\theta(x,x')$ de $(\underset{\star}{X}_k)$ vérifie $0 < \sigma^- \leq \underset{\star}{q}_\theta(x,x') \leq \sigma^+ < +\infty$ pour tous $x, x' \in \underset{\star}{\mathsf{X}}$ et tout $\theta \in \Theta$, et la mesure $\underset{\star}{\lambda}$ est une mesure de probabilité.*
- *Pour tout $y \in \underset{\star}{\mathsf{Y}}$, l'intégrale $\int_{\underset{\star}{\mathsf{X}}} \underset{\star}{g}_\theta(x,y)\underset{\star}{\lambda}(dx)$ vérifie :*

$$\inf_{\theta\in\Theta} \int_{\underset{\star}{\mathsf{X}}} \underset{\star}{g}_\theta(x,y)\underset{\star}{\lambda}(dx) > 0$$

et

$$\sup_{\theta\in\Theta} \int_{\underset{\star}{\mathsf{X}}} \underset{\star}{g}_\theta(x,y)\underset{\star}{\lambda}(dx) < +\infty.$$

Hypothèse 7.1.6. $b^+ = \sup_\theta \sup_{x,y} \underset{\star}{g}_\theta(x,y) < +\infty$ *et* $E_{\theta_0}|\log b^-(\underset{\star}{Y}_1)| < +\infty$, *où*

$$b^- = \inf_\theta \int_{\underset{\star}{\mathsf{X}}} \underset{\star}{g}_\theta(x,y)\underset{\star}{\lambda}(dx).$$

Hypothèse 7.1.7. *Pour tous $(x,x') \in \underset{\star}{\mathsf{X}} \times \underset{\star}{\mathsf{X}}$ et $y \in \underset{\star}{\mathsf{Y}}$, les fonctions $\theta \mapsto \underset{\star}{q}_\theta(x,x')$ et $\theta \mapsto \underset{\star}{g}_\theta(x,y)$ sont continues.*

Hypothèse 7.1.8. *Il existe un voisinage ouvert $\mathcal{U} = \{\theta; |\theta - \theta_0| < \delta\}$ de θ_0 vérifiant :*

- *Pour tout $(x,x') \in \underset{\star}{\mathsf{X}} \times \underset{\star}{\mathsf{X}}$ et tout $y \in \underset{\star}{\mathsf{Y}}$, les fonctions $\theta \mapsto \underset{\star}{q}_\theta(x,x')$ et $\theta \mapsto \underset{\star}{g}_\theta(x,y)$ sont deux fois continûment différentiables sur $\mathcal{U}$.*
- $\sup_{\theta\in\mathcal{U}} \sup_{x,x'} \|\nabla_\theta \log \underset{\star}{q}_\theta(x,x')\| < +\infty \quad et \quad \sup_{\theta\in\mathcal{U}} \sup_{x,x'} \|\nabla^2_\theta \log \underset{\star}{q}_\theta(x,x')\| < +\infty.$
- $E_{\theta_0}\left[\sup_{\theta\in\mathcal{U}} \sup_x \|\nabla_\theta \log \underset{\star}{g}_\theta(x,\underset{\star}{Y}_1)\|^2\right]$ *et* $E_{\theta_0}\left[\sup_{\theta\in\mathcal{U}} \sup_x \|\nabla^2_\theta \log \underset{\star}{g}_\theta(x,\underset{\star}{Y}_1)\|\right]$ *sont finies.*
- *Pour $\underset{\star}{\mu}$-presque tout $y \in \underset{\star}{\mathsf{Y}}$, il existe une fonction $f_y : \underset{\star}{\mathsf{X}} \to \mathbb{R}_+$ de $L^1(\underset{\star}{\lambda})$ telle que $\sup_{\theta\in\mathcal{U}} \underset{\star}{g}_\theta(x,y) \leq f_y(x)$.*
- *Pour $\underset{\star}{\lambda}$-presque tout $x \in \underset{\star}{\mathsf{X}}$, il existe une fonction $f^1_x : \underset{\star}{\mathsf{Y}} \to \mathbb{R}_+$ et $f^2_x : \underset{\star}{\mathsf{Y}} \to \mathbb{R}_+$ de $L^1(\underset{\star}{\mu})$ telles que $\|\nabla_\theta \underset{\star}{g}_\theta(x,y)\| \leq f^1_x(y)$ et $\|\nabla^2_\theta \underset{\star}{g}_\theta(x,y)\| \leq f^2_x(y)$ pour tout $\theta \in \mathcal{U}$.*

On suppose que la chaîne de Markov $(\underset{\star}{X}_i)_{i\in 1:m}$ est initialisée en $\underset{\star}{X}_1 = x_1$. On définit la log-vraisemblance par :

$$l_{x_1,m}(\theta) = \sum_{i=1}^{m} \log\left(\int g_\theta(x_i, \underset{\star}{Y}_i) P_\theta(\underset{\star}{X}_i \in dx_i \mid \underset{\star}{Y}_{1:i-1}, \underset{\star}{X}_1 = x_1)\right)$$

et $\hat{\theta}_{x_1,m}$ l'estimateur du maximum de vraisemblance associé (correspondant à un argument du maximum de la fonction $l_{x_1,m}(.)$). On a alors le théorème suivant :

Théorème 7.1.9. *On suppose que les cinq hypothèses précédentes sont vérifiées et que de plus θ_0 est identifiable. Alors les points suivants sont vérifiés.*

- *L'estimateur du maximum de vraisemblance $\hat{\theta}_m = \hat{\theta}_{x_1,m}$ est consistant :*

$$\hat{\theta}_m \to_{m\to+\infty} \theta_0 \ P_{\theta_0}\text{-}p.s..$$

- *Si la matrice d'information de Fisher $\mathcal{I}(\theta_0)$ est inversible et θ_0 est dans l'intérieur de Θ, alors l'estimateur du maximum de vraisemblance est asymptotiquement normal :*

$$m^{1/2}(\hat{\theta}_m - \theta_0) \to_{m\to+\infty} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1}) \text{ en loi (par rapport à } P_{\theta_0})$$

- *La matrice d'information observée en l'estimateur du maximum de vraisemblance est un estimateur consistant de $\mathcal{I}(\theta_0)$:*

$$-n^{-1}\nabla^2_\theta l_{x_1,m}(\hat{\theta}_m) \to_{m\to+\infty} \mathcal{I}(\theta_0) \ P_{\theta_0}\text{-}p.s..$$

Démonstration. Chapitre 12 de [21] jusqu'à la section 12.5.5. Le plan de la preuve suit celle de la consistance du maximum de vraisemblance dans le cas indépendant de Wald [116], adapté au comportement du la chaîne de Markov cachée. □

7.2 Description du noyau de transition

Le but de cette section suivante est de définir explicitement l'espace d'états des variables Z_i ainsi qu'une probabilité de référence λ associée telle que le noyau de transition Q de Z_i vers Z_{i+1} vérifie : $Q(z, .) \ll \lambda$ pour toute évolution z.

On considère dans les sections 7.2.1 et 7.2.2 que l'évolution s'effectue de séquence à séquence sur l'intervalle $[0, T]$. On adaptera dans la section 7.2.3 ces constructions dans le cas d'un arbre enraciné.

7.2.1 Mesure de référence

Introduction. Jusqu'à maintenant, l'espace d'états Z des variables Z_i n'a pas été explicitement construit. On souhaite écrire $\mathsf{Z} = \mathsf{G} \times \mathsf{H}$ avec G l'ensemble des évolutions du premier nucléotide ρ-encodé et H l'ensemble des évolutions du deuxième nucléotide η-encodé.

Néanmoins, la construction de G, de sa tribu et de sa mesure associée n'est pas évidente. En effet, supposons que l'on choisisse p l'évolution π-encodée définie par le nucléotide R sur $[0, T/2[$ et le nucléotide Y sur $[T/2, T]$. On choisit ensuite $\mathsf{B} \subset \mathsf{G}$ l'ensemble des évolutions ayant leur évolution π-encodée associée fixée par p (ensemble que l'on souhaite mesurable).

Si on utilise la mesure de Lebesgue pour chaque saut dans $\{R, C, T\}$ pour construire une mesure sur G, alors la mesure de l'ensemble B est nulle et $\lambda(\mathsf{B} \times \mathsf{H}) = 0$. Cela n'est

pas convenable puisque d'autre part, on a $Q(z, \mathsf{B}) \neq 0$ pour certaines évolutions z (explicitement : pour les évolutions telles que l'évolution du deuxième nucléotide π-encodé soit égal à p).

Pour contourner ce problème, on va utiliser une construction de G où la mesure de Lebesgue est utilisée uniquement pour les sauts de C vers T et de T vers C.

On veut alors décrire une évolution z de la façon suivante.

Mémento 7.2.1. *Une évolution $z \in \mathsf{Z}$ peut se résumer en :*

$$z = (k, (g_0, \dots g_k), j, (s_1, \dots s_j), l, (h_0, \dots h_l), (r_0 \dots r_l)) \,,$$

où :

- *k est le nombre de changements au site gauche dans $\{R, C, T\}$ et $(g_0, \dots g_k)$ les valeurs de changements.*
- *j est le nombre de changements au site gauche de C vers T ou de T vers C et $(s_1, \dots s_j)$ les instants de changements.*
- *l est le nombre de changements au site droit dans $\{A, G, Y\}$, $(h_0, \dots h_l)$ les valeurs de changements, pour les instants $(0 = r_0 \dots r_l)$.*

On note aussi lorsque l'on veut faire apparaître le rattachement à z :

$$z = (k(z), g(z), j(z), s(z), l(z), h(z), r(z)) \,.$$

Par conséquent, lorsque l'on écrit $Z_{i-1} = z$ et $Z_i = z'$ (pour un site i), on peut reconstituer l'ensemble des changements et les instants de sauts survenus au site i dans $\{R, C, T\}$ (grâce à $l(z)$, $h(z)$, $r(z)$, $k(z')$, $g(z')$, $j(z')$ et $s(z')$) et au site $i+1$ dans $\{A, G, Y\}$ (grâce à $l(z')$, $h(z')$ et $r(z')$).

Exemple 7.2.2. *On choisit $z \in \mathsf{Z}$ tel que $k = 6$, $(g_0, \dots g_k) = (R, C, T, C, T, R, C)$, $j = 3$, $l = 5$, $(h_0, \dots h_l) = (A, G, Y, A, G, A)$, et des instants $(s_1, \dots s_j)$ et $(0 = r_0 \dots r_l)$. On peut alors représenter z par la figure 7.1.*

Les points d'interrogation correspondent à des dates qui ne sont pas connues à travers z.

Passons maintenant à la construction effective de $\mathsf{Z} = \mathsf{G} \times \mathsf{H}$. On utilise dans cette section le symbole $\coprod$ pour désigner l'union disjointe des ensembles.

Mesure sur les simplexes. Comme dans la section 1.5.2, on pose pour $l \in \mathbb{N}$: $\mathsf{F}_l := \{(t_0, \dots, t_l); 0 = t_0 < t_1 < \dots < t_l < T\}$. Pour $l \geq 1$, chaque élément de F_l est vu comme un élément $(t_1, \dots, t_l)$ de $\mathbb{R}^l$ et on définit μ_l la mesure de Lebesgue sur la tribu borélienne $\mathcal{F}_l$ associée à F_l. Pour $l = 0$, on définit μ_0 par $\mu_0(\mathsf{F}_0) := 1$.

Mesure pour le deuxième nucléotide. On pose pour $l \in \mathbb{N}$:

$$\mathsf{H}_l := \{A, G, Y\}^{l+1} \times \mathsf{F}_l \text{ et } \mathsf{H} := \coprod_{l \in \mathbb{N}} \mathsf{H}_l.$$

On associe à $\{A, G, Y\}^{l+1}$ la mesure discrète $\delta_{2,l}$ suivante ; pour $n_1, \dots, n_{l+1} \in \{A, G, Y\}$:

s	i	i+1	r
	R	A	r_0
?	C		
s_1	T		
		G	r_1
		Y	r_2
s_2	C		
		A	r_3
s_j	T		
?	R		
		G	r_4
?	C		
		A	r_l
	C	A	T

FIGURE 7.1 – Évolution z de l'exemple 7.2.2.

– $\delta_{2,l}(n_1, \ldots, n_{l+1}) := 1$ si pour tout $u \in [\![1, l]\!]$, $n_u \neq n_{u+1}$.
– $\delta_{2,l}(n_1, \ldots, n_{l+1}) := 0$ sinon.

Pour chaque H_l on lui associe sa tribu produit $\mathcal{H}_l$ et la mesure $\nu_{2,l}$ produit de $\delta_{2,l}$ et de μ_l. On pose ensuite $\mathcal{H} := \sigma\left(\coprod_{l\in\mathbb{N}} \mathcal{H}_l\right)$. Pour $\mathsf{B} \in \mathcal{H}$, il existe une famille $\mathsf{B}_l \in \mathcal{H}_l$, $l \in \mathbb{N}$, tels que $\mathsf{B} = \coprod_{l\in\mathbb{N}} \mathsf{B}_l$, et on définit la mesure ν_2 par :

$$\nu_2(\mathsf{B}) := \sum_{l\in\mathbb{N}} \nu_{2,l}(\mathsf{B}_l).$$

ν_2 est alors une mesure finie sur $(\mathsf{H}, \mathcal{H})$ de masse plus petite que $3e^{3T}$ (car la masse de $\delta_{2,l}$ est plus petite que 3^{l+1} et celle de μ_l est $T^l/l!$).

Mesure pour le premier nucléotide. On pose pour $k \in \mathbb{N}$:

$$\mathsf{G}_k := \coprod_{j=0}^{k} \left(\{R, C, T\}^{k+1} \times \mathsf{F}_j\right) \quad \text{et} \quad \mathsf{G} := \coprod_{k\in\mathbb{N}} \mathsf{G}_k.$$

Soit $k \in \mathbb{N}$ et $j \in [\![0, k]\!]$. On définit la mesure $\mu_{k,j}$ sur $(\mathsf{F}_j, \mathcal{F}_j)$ telle que pour $\mathsf{B} \in \mathcal{F}_j$:

$$\mu_{k,j}(\mathsf{B}) := \frac{T^{k-j}}{(k-j)!}\mu_j(\mathsf{B}).$$

On remarque ici que $\mu_{k,j}(\mathsf{F}_j) = \binom{k}{j}\frac{T^k}{k!} \leq \frac{(2T)^k}{k!}$.
On associe à $\{R, C, T\}^{k+1}$ la mesure discrète $\delta_{1,k,j}$ suivante ; pour $g_1, \ldots, g_{k+1} \in \{R, C, T\}$:

– $\delta_{1,k,j}(g_1,\dots,g_{k+1}) := 1$ si pour tout $u \in [\![1,k]\!]$, $g_u \neq g_{u+1}$ et s'il existe exactement j changements entre C et T (c'est-à-dire s'il existe exactement j entiers $(c_1,\dots,c_j) \subset [\![1,k]\!]$ tels que pour chaque $u' \in (c_1,\dots,c_j)$, ($g_{u'} = C$ et $g_{u'+1} = T$) ou ($g_{u'} = T$ et $g_{u'+1} = C$)).
– $\delta_{1,k,j}(g_1,\dots,g_{k+1}) := 0$ sinon.

On associe à l'ensemble $\{R,C,T\}^{k+1} \times \mathsf{F}_j$ la tribu produit associée et la mesure $\nu_{1,k,j}$ produit de $\delta_{1,k,j}$ et de $\mu_{k,j}$.

On en déduit une tribu et une mesure pour G_k puis pour G (de la même façon que pour le deuxième nucléotide). On a ainsi une mesure finie notée ν_1 sur $(\mathsf{G},\mathcal{G})$, de masse plus petite que $3e^{6T}$.

Définition de Z et mesure associée.

Définition 7.2.3. *Pour obtenir des mesures de probabilités, on normalise ν_1 (resp. ν_2) en ν_1^* (resp. ν_2^*). En tout, on pose $\mathsf{Z} := \mathsf{G} \times \mathsf{H}$, muni de sa tribu produit $\mathcal{Z}$ et de la mesure λ produit de ν_1^* et ν_2^*.*

7.2.2 Noyau de transition

Transition compatible.

Définition 7.2.4. *Pour $m = (m_0,\dots m_n)$ une suite finie à valeurs dans $\{A,C,G,T,R,Y\}$, on définit la séquence $\tilde{\pi}(m)$ la séquence associée définie par l'algorithme suivant :*
– *Débuter avec m.*
– *Remplacer chaque A et G par R. Remplacer chaque C et T par Y.*
– *S'il y a plus de deux R consécutifs ou Y consécutifs, supprimer les doublons.*

Exemple 7.2.5. *(A,G,Y,G,Y,A,G,A,G,Y) devient (R,Y,R,Y,R,Y).*

On définit aussi I_z la réunion des intervalles où le deuxième nucléotide dans z est égal à Y.

Définition 7.2.6. *Pour $z = (k,(g_0,\dots g_k),j,(s_1,\dots s_j),l,(h_0,\dots h_l),(r_0\dots r_l))$ et en notant r_{l+1} l'instant final, on pose :*

$$I_z = \{r \in]r_d, r_{d+1}[\ ;\ d \text{ tel que } h_d = Y\}.$$

Définition 7.2.7. *Soit :*

$$\begin{aligned} z &= (k(z),g(z),j(z),s(z),l(z),h(z),r(z))\ , \\ z' &= \big(k(z'),g(z'),j(z'),s(z'),l(z'),h(z'),r(z')\big) \in \mathsf{Z}. \end{aligned}$$

On dit que z est compatible avec z' et on écrit $z \rightsquigarrow z'$ si :
– $\tilde{\pi}(h(z)) = \tilde{\pi}(g(z'))$
– $s(z') \subset I_z$

Écriture de la transition.

Remarque 7.2.8. *Dans ce paragraphe seulement, on distingue explicitement $Z_i = (\rho_i, \eta_{i+1})$ et $\tilde{Z}_i$ la projection qui à une évolution z issue de Z_i est associé l'élément de Z correspondant. La projection correspond à oublier les instants de sauts entre R et $\{C, T\}$ au site i.*

On fixe un site i et on suppose que $\tilde{Z}_{i-1} = z$, $\tilde{Z}_i = z'$ et $z \rightsquigarrow z'$. D'après le mémento 7.2.1, on peut définir tous les temps de changements $(t_l)_{l\in[\![1,L]\!]}$ survenant dans $\{C, T, R\} \times \{A, G, Y\}$ aux sites $(i, i+1)$, et avec $t_0 = 0$, on définit pour tout $l \in [\![1, L]\!]$: $\Delta_l = t_l - t_{l-1}$. On associe alors à chaque Δ_l la valeur a_l parmi $\{C, T, R\} \times \{A, G, Y\}$ prise par l'évolution sur l'intervalle $[t_{l-1}, t_l[$.

On reprend les notations du théorème 6.2.3 et on pose :

$$W(a_l, a_l) = \begin{cases} W_R(a_l, a_l) & \text{si } a_l \in \{RA, RG, RY\}, \\ W_Y(a_l, a_l) & \text{sinon,} \end{cases}$$

et

$$V(a_l, a_{l+1}) = \begin{cases} W(a_l, a_{l+1}) & \text{si } a_l, a_{l+1} \in \{RA, RG, RY\} \text{ ou } a_l, a_{l+1} \in \{CA, CG, CY, TA, TG, TY\}, \\ U_{R\to Y}(a_l, a_{l+1}) & \text{si } a_l \in \{RA, RG, RY\} \text{ et } a_{l+1} \in \{CA, CG, CY, TA, TG, TY\}, \\ U_{Y\to R}(a_l, a_{l+1}) & \text{sinon.} \end{cases}$$

D'après ce même théorème 6.2.3 on aboutit à la densité de transition suivante :

Proposition 7.2.9. *Pour $z \rightsquigarrow z'$ on note,*

$$\check{q}(z, z') = \left(\prod_{l=1}^{L-1} e^{\Delta_l W(a_l, a_l)} V(a_l, a_{l+1}) \right) e^{\Delta_L W(a_L, a_L)}.$$

La densité de transition de $\tilde{Z}_{i-1} = z$ vers $\tilde{Z}_i = z'$ vérifie si $z \rightsquigarrow z'$:

$$q(z, z') = \frac{1}{C(z)} \check{q}(z, z')$$

avec $C(z)$ coefficient de normalisation donné par :

$$C(z) = \int_{\mathsf{Z}} \check{q}(z, z') d\lambda(z').$$

Lorsque, $z \not\rightsquigarrow z'$, on a $q(z, z') = 0$.

On peut en outre identifier en chaque site les variables $\tilde{Z}_i$ (d'espace d'états Z) et $Z_i = (\rho_i, \eta_{i+1})$. De même, on identifie ρ_i avec la projection de $\tilde{Z}_i$ sur le premier nucléotide (l'espace d'états associé est G) et ρ_{i+1} avec la projection de $\tilde{Z}_i$ sur le deuxième nucléotide (l'espace d'états associé est H).

On identifie également l'espace d'états de la variable X_i avec $\mathsf{H} \times \mathsf{G}$.

De plus, on définit la probabilité de transition pour tous $z \in \mathsf{Z}$ et $\mathsf{B} \in \mathcal{Z}$ par

$$Q(z, \mathsf{B}) := \int_{\mathsf{B}} q(z, z') d\lambda(z').$$

Démonstration. Soit $z \in \mathsf{Z}$. On vérifie d'abord que l'intégrale $\int_{\mathsf{Z}} \check{q}(z,z')d\lambda(z')$ est finie. Comme tous les paramètres de taux de sauts sont finis, il existe $M \geq 0$ tel que pour tout $l \in [\![1, L-1]\!]$,

$$V(a_l, a_{l+1}) \leq M.$$

Comme de plus pour tout $l \in [\![1, L-1]\!]$, $W(a_l, a_l) \leq 0$, on obtient :

$$\check{q}(z,z') \leq M^{L-1}.$$

Or, $L-1$ correspond au nombre de changements $l(z') + k(z')$ dans $\{C,T,R\} \times \{A,G,Y\}$ aux sites encodés $(i, i+1)$. D'après la définition 7.2.3, on écrit $z' = (\mathsf{g}, \mathsf{h}) \in \mathsf{G} \times \mathsf{H}$ et on remarque que $k(z') = k(\mathsf{g})$ et $l(z') = l(\mathsf{h})$. De plus, on a :

$$\int_{\mathsf{G}} M^{k(\mathsf{g})} \nu_1(\mathsf{g}) \leq \sum_{k \geq 0} M^k \frac{3^{k+1}(2T)^k}{k!} = 3e^{6MT} < +\infty$$

et

$$\int_{\mathsf{H}} M^{l(\mathsf{h})} \nu_2(\mathsf{h}) \leq \sum_{l \geq 0} M^l \frac{3^{l+1} T^l}{l!} = 3e^{3MT} < +\infty.$$

On en déduit alors que $\int_{\mathsf{Z}} \check{q}(z,z')d\lambda(z') < +\infty$.

D'autre part, connaître $\tilde{Z}_{i-1}$ correspond à connaître l'évolution $\eta(i)$, et une partie de ρ_{i-1}. De plus, connaître $(\tilde{Z}_{i-1}, \tilde{Z}_i)$ correspond à connaître les évolutions $\eta(i)$ et Z_i, et une partie de ρ_{i-1}. Ainsi avec l'encodage choisi pour décrire les évolutions $z \in \mathsf{Z}$, le noyau de transition de z vers z' coïncide bien avec celui issu de la matrice de transition décrit dans le théorème 6.2.3. □

7.2.3 Écriture sur un arbre

On a considéré jusqu'ici l'évolution de séquence à séquence. Le long d'un arbre, on utilise une nouvelle fois la construction de l'évolution pour le modèle complet de la définition 1.3.7 (en particulier la propriété d'indépendance le long des arêtes de l'arbre, qui permet ainsi de décrire l'évolution au niveau des branchements de l'arbre) pour décrire l'évolution de façon analogue à celle décrite dans la section 6.4.2.

Pour simplifier l'écriture, une évolution $z \in \mathsf{Z}$ de séquence à séquence décrit également une évolution d'un couple encodé le long d'un arbre. $z(T)$ désigne alors l'ensemble des dinucléotides aux feuilles de l'arbre.

7.2.4 Noyau de transition vers les observations

On choisit un arbre T comportant $f(\mathsf{T})$ feuilles. On rappelle que les observations correspondent aux séquences (supposées de longueur m) associées aux feuilles de l'arbre considéré. Pour chaque site $i \in [\![1, m-1]\!]$, on note $Z_i(T)$ l'ensemble des dinucléotides Φ-encodés observés aux feuilles de l'arbre associé à l'évolution Z_i (le long de l'arbre).

On décrit les observations $y := Z_i(T)$ de la façon suivante :

Définition 7.2.10. *On pose :*

$$\mathsf{Z}(T) = \{CA, CG, CY, TA, TG, TY, RA, RG, RY\}^{f(\mathsf{T})},$$

muni de la tribu discrète $\mathcal{Z}(T)$ et de la mesure μ discrète. On pose aussi pour tous $z \in \mathsf{Z}$, $y \in \mathsf{Z}(T)$:

$$g(z, y) = \mathbf{1}(z(T) = y).$$

On définit enfin pour tous $z \in \mathsf{Z}$, $\mathsf{B} \in \mathcal{Z}(T)$ la probabilité de transition : $G(z, \mathsf{B}) = \int_B g(z, y) d\mu(y)$.

7.3 Adaptation du théorème pour notre modèle

On va étudier la validité des hypothèses 7.1.4 à 7.1.8 permettant d'appliquer le théorème 7.1.9, dans le cas de notre chaîne de Markov cachée. Cette chaîne de Markov cachée correspond au modèle associé aux espaces $(\mathsf{Z}, \mathcal{Z})$ et $(\mathsf{Z}(\mathsf{T}), \mathcal{Z}(T))$, au noyau de transition Q sur $(\mathsf{Z}, \mathcal{Z})$ et au noyau de transition G de $(\mathsf{Z}, \mathcal{Z})$ vers $(\mathsf{Z}(\mathsf{T}), \mathcal{Z}(T))$ (ces espaces et noyaux ont été définis dans la section 7.2).

Tout d'abord, on impose les conditions de compacité et d'identifiabilité suivantes de l'ensemble de modèles Θ considéré. On suppose que tous les modèles globaux RN95+YpR considérés possèdent la même loi à la racine R_0 et que la topologie de l'arbre et les différentes longueurs de branches T_0 sont fixées.

De plus, les taux de sauts d'un modèle RN95+YpR peuvent être décrits à l'aide d'un 16-uplets $(v_x, w_x, r_y; x \in \mathcal{A}, y \in \mathcal{B})$, avec pour tous $x \in \mathcal{A}$ et $y \in \mathcal{B}$, $v_x > 0$, $w_x > 0$ et $r_y \geq 0$ (voir notation 1.2.4).

Ainsi, l'ensemble des paramètres possibles peut être vu comme l'ensemble $\Upsilon = (\mathbb{R}_*^+)^8 \times (\mathbb{R}^+)^8$. Dans tout ce chapitre, on fait l'hypothèse que l'espace des paramètres est compact.

Hypothèse 7.3.1. *L'ensemble des paramètres Θ considérés est un sous-espace compact de Υ.*

On fait également l'hypothèse suivante d'identifiabilité des modèles.

Hypothèse 7.3.2. *Les modèles RN95+YpR associés aux éléments de l'ensemble Θ sont identifiables.*

L'ensemble des modèles considérés s'écrit donc de la forme $(\mathsf{R}_0, \mathsf{T}_0, \mathsf{M})$, avec M régie par un paramètre $\theta \in \Theta$.

On verra que toutes les hypothèses 7.1.4 à 7.1.8 sont vérifiées hormis l'hypothèse 7.1.5. Cette hypothèse sert en particulier à vérifier la propriété d'oubli de la condition initiale au fur et à mesure que le nombre de sites devient important. On montrera que dans notre cas, cette propriété d'oubli est encore vérifiée ce qui permettra d'adapter la preuve du théorème 7.1.9 pour notre modèle.

7.3.1 Condition de Doeblin du noyau Q

Pour montrer la validité de l'hypothèse 7.1.4, on vérifie une condition de Doeblin du noyau Q.

Les conditions de Doeblin sont définies dans la section 4.3.3 de [21]. On montre ici une condition de Doeblin globale sur Q^2 :

Proposition 7.3.3. *Q^2 vérifie une condition de Doeblin globale, c'est-à-dire il existe $\varepsilon > 0$ et une mesure de probabilité ϑ sur $(\mathsf{Z}, \mathcal{Z})$ telle que pour tout $z \in \mathsf{Z}$ et $D \in \mathcal{Z}$, on ait :*

$$Q^2(z, D) \geq \varepsilon \vartheta(D).$$

En outre, il existe une unique probabilité invariante (voir par exemple le théorème 4.3.16 de [21]).

Pour montrer cette proposition, nous allons tout d'abord exprimer la densité de η_{i+2} conditionnellement à une évolution $X_{i+1} = x_{i+1}$ (l'évolution au cours du temps au site $i+1$). Ensuite, en utilisant le fait que les coefficients de taux de sauts v_x et w_x (pour $x \in \mathcal{A}$) sont strictement positifs d'après l'hypothèse 7.3.1 sur l'ensemble des paramètres Θ, on en déduit une minoration de cette densité.

On peut alors en déduire $\varepsilon > 0$ et une mesure de probabilité ϑ_0 telle que pour $C \in \mathcal{H}$ et tout $z \in \mathsf{Z}$, on ait :

$$P(\eta_{i+2} \in C | Z_i = z) \geq \varepsilon \vartheta_0(C).$$

De cette inégalité on en déduit enfin une condition de Doeblin globale pour Q^2.

Démonstration. On considère la construction de la mesure de référence $\lambda = \nu_1^* \otimes \nu_2^*$ de la section 7.2 sur l'espace $\mathsf{Z} := \mathsf{G} \times \mathsf{H}$ et de la proposition 7.2.9. On choisit un site i. On souhaite exprimer la densité conditionnelle de η_{i+2} conditionnellement à X_{i+1}. D'après la définition 3.2.6 des matrices de taux de sauts $Q_{g,}$ pour $g \in \{R, C, T\}$,

$$Q_{g,} = \begin{array}{c} \\ A \\ G \\ Y \end{array} \begin{array}{c} \begin{array}{ccc} A & G & Y \end{array} \\ \begin{pmatrix} . & \begin{array}{c} w_G + \\ r_{TA\to TG}\mathbf{1}_{g=T} + \\ r_{CA\to CG}\mathbf{1}_{g=C} \end{array} & v_T + v_C \\ \begin{array}{c} w_A + \\ r_{TG\to TA}\mathbf{1}_{g=T} + \\ r_{CG\to CA}\mathbf{1}_{g=C} \end{array} & . & v_T + v_C \\ v_A & v_G & . \end{pmatrix} \end{array},$$

et en utilisant le théorème 3.2.10, on exprime pour pour $x = (\mathsf{h}_0, \mathsf{g}) \in \mathsf{H} \times \mathsf{G}$ la densité conditionnelle de η_{i+2} conditionnellement à $X_{i+1} = x$ par la fonction f suivante, pour $\mathsf{h} \in \mathsf{H}$:

$$f(\eta_{i+2} = \mathsf{h} | X_{i+1} = x) = \prod_{l=1}^{L} e^{\Delta_l Q_{g_l,}(a_l, a_l)} \prod_{l=1}^{L-1} Q_{g_l,}(a_l, a_{l+1}),$$

avec (Δ_l) les durées de changements dans l'ensemble $\{R, C, T\} \times \{A, G, R\}$ et les valeurs associées (g_l) au site $i+1$ et (a_l) au site $i+2$.

On observe que pour tout $g \in \{R, C, T\}$, on a les égalités et inégalités suivantes :

$$\begin{aligned} w_G \leq & Q_{g,}(A, G) \leq w_G + \max(r_{CA\to CG}, r_{TA\to TG}) \\ w_A \leq & Q_{g,}(G, A) \leq w_A + \max(r_{CG\to CA}, r_{TG\to TA}) \\ & Q_{g,}(A, Y) = v_C + v_T \\ & Q_{g,}(G, Y) = v_C + v_T \\ & Q_{g,}(Y, A) = v_A \\ & Q_{g,}(Y, G) = v_G. \end{aligned}$$

D'après l'hypothèse 7.3.1 sur l'ensemble des paramètres Θ, il existe des constantes c_1 et c_2 strictement positives telle que les termes $Q_{g_l,}(a_l, a_{l+1})$ soient minorés par c_1 et les termes $Q_{g_l,}(a_l, a_l)$ soient minorés par $-c_2$. Cela conduit à :

$$f(\eta_{i+2} = \mathsf{h} | X_{i+1} = x) \geq \prod_{l=1}^{L} e^{-\Delta_l c_2} c_1^{L-1} = e^{-Tc_2} c_1^{L-1}.$$

Comme dans la démonstration de la proposition 7.2.9, on écrit le nombre total de sauts $L - 1 = k(\mathsf{g}) + l(\mathsf{h})$, donc :

$$f(\eta_{i+2} = \mathsf{h} | X_{i+1} = x) \geq \left(e^{-Tc_2} c_1^{k(\mathsf{g})}\right) c_1^{l(\mathsf{h})}. \tag{7.2}$$

On écrit ensuite pour $z = (\mathsf{g}_0, \mathsf{h}_0)$ et $C \in \mathcal{H}$:

$$\begin{aligned} P(\eta_{i+2} \in C | Z_i = z) &= \int_C f(\eta_{i+2} = \mathsf{h} | Z_i = z) d\nu_2^*(\mathsf{h}) \\ &= \int_C \int_{\mathsf{G}} f(\rho_{i+1} = \mathsf{g}, \eta_{i+2} = \mathsf{h} | Z_i = z) d\nu_1^*(\mathsf{g}) d\nu_2^*(\mathsf{h}) \\ &= \int_C \int_{\mathsf{G}} f(\eta_{i+2} = \mathsf{h} | X_{i+1} = (\mathsf{h}_0, \mathsf{g})) f(\rho_{i+1} = \mathsf{g} | Z_i = z) d\nu_1^*(\mathsf{g}) d\nu_2^*(\mathsf{h}) \\ &= \int_{\mathsf{G}} f(\rho_{i+1} = \mathsf{g} | Z_i = z) \left[\int_C f(\eta_{i+2} = \mathsf{h} | X_{i+1} = (\mathsf{h}_0, \mathsf{g})) d\nu_2^*(\mathsf{h}) \right] d\nu_1^*(\mathsf{g}). \end{aligned}$$

En utilisant l'équation (7.2), on obtient :

$$P(\eta_{i+2} \in C | Z_i = z) \geq e^{-Tc_2} \int_{\mathsf{G}} f(\rho_{i+1} = \mathsf{g} | Z_i = z) c_1^{k(\mathsf{g})} \left[\int_C c_1^{l(\mathsf{h})} d\nu_2^*(\mathsf{h}) \right] d\nu_1^*(\mathsf{g}).$$

De la même manière que dans la démonstration de la proposition 7.2.9 pour montrer la finitude des intégrales, on montre qu'il existe $c_3 > 0$ et une mesure de probabilité ϑ_0 telle que pour tous $C \in \mathcal{H}$

$$\vartheta_0(C) = c_3 \left[\int_C c_1^{l(\mathsf{h})} d\nu_2^*(\mathsf{h}) \right]$$

On obtient alors l'existence d'un réel strictement positif ε tel que :

$$P(\eta_{i+2} \in C | Z_i = z) \geq e^{-Tc_2} c_3 \int_{\mathsf{G}} f(\rho_{i+1} = \mathsf{g} | Z_i = z) c_1^{k(\mathsf{g})} d\nu_1^*(\mathsf{g}) \vartheta_0(C) = \varepsilon \vartheta_0(C).$$

On a alors pour tous $D \in \mathcal{Z}$ et $z \in \mathsf{Z}$:

$$\begin{aligned} P(Z_{i+2} \in D \mid Z_i = z) &= \int_{\mathcal{H}} P(Z_{i+2} \in D \mid \eta_{i+2} = \mathsf{h}) dP(\eta_{i+2} \in \mathsf{h} \mid Z_i = z) \\ &\geq \varepsilon \int_H P(Z_{i+2} \in D \mid \eta_{i+2} = \mathsf{h}) d\vartheta_0(\mathsf{h}) \\ &=: \varepsilon \vartheta(D) \end{aligned}$$

qui fournit la condition de Doeblin souhaitée.

□

7.3.2 Vérification des hypothèses du théorème 7.1.9

Hypothèses vérifiées. On montre que les hypothèses 7.1.4, 7.1.6, 7.1.7 et 7.1.8 sont vérifiées.

Pour l'hypothèse 7.1.4, on utilise les mesures λ et μ des sections 7.2.1 et 7.2.4. La condition de Doeblin de la proposition 7.3.3 montre le troisième point de l'hypothèse (voir par exemple le théorème 4.3.16 dans [21]).

L'hypothèse 7.1.6 est vérifiée puisque $b^+ = 1$, et comme g est indépendante de θ, on a pour tout $y \in \mathsf{Z}(T)$, $b^-(y) = \lambda\{z \in \mathsf{Z}; g(z, y) \neq 0\} > 0$.

Les hypothèses 7.1.7 et 7.1.8 sont vérifiées par définition de q_θ (voir proposition 7.2.9) et en utilisant que g_θ ne dépend pas de θ.

Pour l'hypothèse d'identifiabilité, on se réfère à l'annexe B.

Hypothèse non vérifiée. L'hypothèse 7.1.5 n'est pas vraie pour notre modèle, puisque $q(x, x') = 0$ pour tous $x \not\rightsquigarrow x'$ (voir la définition 7.2.7 pour la définition de la relation $\rightsquigarrow$).

Dans la preuve du théorème 7.1.9, cette hypothèse sert à obtenir l'oubli de la condition initiale (choisie pour le premier site) lorsque le nombre de sites considéré augmente. Précisément, on a la proposition suivante (avec les notations de la définition 7.1.1) sous l'hypothèse 7.1.5 :

Proposition 7.3.4. *Il existe $\rho < 1$ tel que pour tout $k \geq 1$, tout $y_{1:m}$ et toutes lois initiales ν et ν' de $(X, \mathcal{X})$:*

$$\left\| \int_X P(X_k \in . \mid X_1 = x, Y_{1:m} = y_{1:m})[\nu(dx) - \nu'(dx)] \right\|_{VT} \leq 2\rho^k.$$

On considère dans la section suivante une hypothèse alternative à l'hypothèse 7.1.5 impliquant tout de même (accompagnée des quatre autres hypothèses) le théorème 7.1.9 de consistance et de normalité asymptotique du maximum de vraisemblance et en particulier la proposition 7.3.4.

Notons que dans d'autres contextes, d'autres modifications de l'hypothèse 7.1.5 permettent également d'obtenir la proposition 7.3.4 et, accompagnée des quatre autres hypothèses, le résultat du théorème 7.1.9. Deux modifications issues de [21] ne sont pas utilisables pour la chaîne de Markov cachée considérée ici.

Une première modification consiste à utiliser l'hypothèse 4.3.29 de [21], mais elle impose que la fonction de densité g soit à valeurs dans $]0, +\infty[$.

Une deuxième modification consiste à utiliser l'hypothèse 4.3.31 de [21], mais cette hypothèse nécessite l'existence d'un ensemble 1-*small* (défini dans la section 14.2.2.2 de [21]) pour le noyau Q. Ce n'est pas le cas ici puisque pour tout ensemble $C \in \mathcal{Z}$ de mesure positive et tout $z \in C$, il existe z' vérifiant $q(z, z') = 0$.

7.3.3 Hypothèse alternative

On va énoncer l'hypothèse suivante 7.3.7, vérifiée dans notre modèle, qui s'inspire des techniques utilisées dans la section 4.3.6 de [21]. Cette hypothèse impose que le noyau de transition sur l'espace caché vérifie une condition de Doeblin, et qu'en plus cette condition de Doeblin reste vérifiée quel que soit les observations possibles.

On pose tout d'abord les définitions suivantes.

Définition 7.3.5. *Pour $x, x' \in \underset{\star}{\mathsf{X}}$ et $y, y' \in \underset{\star}{\mathsf{Y}}$, on écrit à θ fixé :*

1. $x \dashrightarrow x'$ *si* $\underset{\star}{q}_\theta(x, x') \neq 0$,
2. $x \dashrightarrow y$ *si* $\underset{\star}{g}_\theta(x, y) \neq 0$,
3. $y \dashrightarrow y'$ *s'il existe* $x \in \underset{\star}{\mathsf{X}}$ *vérifiant* $x \dashrightarrow y$, $A \in \underset{\star}{\mathcal{X}}$ *de mesure strictement positive tel que pour tout* $x' \in A$, *on ait* $x' \dashrightarrow y'$ *et* $x \dashrightarrow x'$.

On note alors pour $d \in \mathbb{N}^*$, $x_1 \in \mathsf{X}$, $A \in \underset{\star}{\mathcal{X}}$, $y_{2:d+1} \in \underset{\star}{\mathsf{Y}}^d$ et à θ fixé :

$$D_\theta[y_{2:d+1}](x_1, A) = \int_{x_2} \dots \int_{x_d} \int_{x_{d+1} \in A} \prod_{l=2}^{d+1} \underset{\star}{Q}(x_{l-1}, dx_l) \mathbf{1}_{x_l \dashrightarrow y_l}.$$

Remarque 7.3.6. *La relation énoncée dans le point 1 de la définition 7.3.5 coïncide avec la relation $\rightsquigarrow$ définie dans la définition 7.2.7 pour notre modèle.*

Hypothèse 7.3.7.

- *Il existe un entier $d_0 \geq 2$, deux nombres strictement positifs σ^- et σ^+, et une mesure $\underset{\star}{\lambda}$ sur $(\underset{\star}{\mathsf{X}}, \underset{\star}{\mathcal{X}})$ tels que pour tous $x \in \underset{\star}{\mathsf{X}}$, $A \in \underset{\star}{\mathcal{X}}$ et $\theta \in \Theta$,*

$$\sigma^- \lambda(A) \leq \underset{\star}{Q}_\theta^{d_0}(x, A) \leq \sigma^+ \lambda(A).$$

- *Il existe deux constantes g^- et g^+ de $]0, +\infty[$ telles que pour tous $y \in \underset{\star}{\mathsf{Y}}$, $\theta \in \Theta$,*

$$g^- \leq \inf_{x, x \dashrightarrow y} \underset{\star}{g}_\theta(x, y) \leq \sup_{x, x \dashrightarrow y} \underset{\star}{g}_\theta(x, y) \leq g^+.$$

- *Pour l'entier d_0 du premier point, il existe $\varepsilon > 0$ tel que pour tout $d \geq d_0$, tous $x_1, x \in \underset{\star}{\mathsf{X}}$, tous y_1, y_{d+1} tels que $x_1 \dashrightarrow y_1$ et $x \dashrightarrow y_{d+1}$, tous $y_2, \dots, y_d$ tels que $y_1 \dashrightarrow \dots \dashrightarrow y_{d+1}$ et tout $\theta \in \Theta$ on ait :*

$$D_\theta[y_{2:d+1}](x_1, dx) > \varepsilon \underset{\star}{\lambda}(dx).$$

Vérification de l'hypothèse pour notre modèle. Avec $d_0 = 2$, le premier point est vérifié d'après la proposition 7.3.3 et puisque q_θ est bornée. Le deuxième point est vérifié avec g^- et g^+ égales à 1.

Pour le troisième point, pour $i \geq d_0$, on choisit $(z_1, z) \in \mathsf{Z}^2$. Un couple (y_1, y_{i+1}) vérifiant $z_1 \dashrightarrow y_1$ et $z \dashrightarrow y_{i+1}$ est nécessairement $(z_1(T), z(T))$. On choisit ensuite $y_2, \ldots, y_i$ vérifiant $y_1 \dashrightarrow y_2 \dashrightarrow \ldots \dashrightarrow y_{i+1}$. On calcule par rapport à λ :

$$\begin{aligned} D_\theta[y_{2:i+1}](z_1, dz) =& P_\theta(Z_{i+1} = dz | Z_1 = z_1, Z_{2:i}(T) = y_{2:i}) \\ =& \frac{P(Z_{2:i}(T) = y_{2:i} | Z_{i+1} = z, Z_1 = z_1)}{P(Z_{2:i}(T) = y_{2:i} | Z_1 = z_1)} P_\theta(Z_{i+1} = dz | Z_1 = z_1) \end{aligned}$$

Ensuite par la proposition 7.3.3, on obtient qu'il existe un réel $\varepsilon_0 > 0$ tel que la minoration suivante soit vérifiée :

$$D_\theta[y_{2:i+1}](z_1, dz) > \varepsilon_0 \frac{P(Z_{2:i}(T) = y_{2:i} | Z_{i+1} = z, Z_1 = z_1)}{P(Z_{2:i}(T) = y_{2:i} | Z_1 = z_1)} \lambda(dz).$$

En majorant $P(Z_{2:i}(T) = y_{2:i} | Z_1 = z_1)$ par 1, on écrit :

$$D_\theta[y_{2:i+1}](z_1, dz) > \varepsilon_0 P(Z_{2:i}(T) = y_{2:i} | Z_{i+1} = z, Z_1 = z_1) \lambda(dz).$$

Or, d'après le théorème 3.2.10, on a :

$$P(Z_{2:i}(T) = y_{2:i} | Z_{i+1} = z, Z_1 = z_1) = P(Z_{2:i}(T) = y_{2:i} | \eta(X_2) = \eta_2, \rho(X_{i+1}) = \rho_{i+1}).$$

En séparant les quatre cas suivants les valeurs de $\pi_2(T)$ et de $\pi_{i+1}(T)$, puis en utilisant l'hypothèse 7.3.1 sur l'ensemble des paramètres Θ et la forme de la densité de l'évolution, on obtient l'existence de $\varepsilon_1 > 0$ vérifiant :

$$P(Z_{2:i}(T) = y_{2:i} | Z_{i+1} = z, Z_1 = z_1) > \varepsilon_1.$$

En choisissant $\varepsilon = \varepsilon_0 \varepsilon_1$, on en déduit le troisième point de l'hypothèse.

Exemple de chaîne de Markov cachée où l'hypothèse n'est pas vérifiée. On prend l'exemple suivant de chaîne de Markov cachée qui est ergodique mais n'admet pas d'oubli de la condition initiale, issu de [21, 25, 69] (respectivement sections 4.3.6, 5 et 10). On pose : $\mathsf{X} = \{0, 1, 2, 3\}$, $\{U_k\}_{k \geq 1}$ loi de Bernoulli de paramètre $1/2$ et on définit $X_1 \in \mathsf{X}$ et pour $k \geq 2$, $X_k = (X_{k-1} + U_k) \mod 4$.

Les observations sont définies par : $Y_k = \mathbf{1}_{\{0,2\}}(X_k)$.

Les deux premiers points de l'hypothèse sont vérifiés avec $d_0 = 4$, mais pour le troisième, en prenant $d = d_0$ et : $(x_1, x_5) = (0, 0)$, $(y_1, y_5) = (0, 0)$, $(y_2, y_3, y_4) = (1, 0, 0)$, on ne peut pas construire (x_2, x_3, x_4) vérifiant la condition voulue.

L'hypothèse implique l'oubli de la condition initiale.

Proposition 7.3.8. *Si l'hypothèse 7.3.7 est vraie alors il existe $\rho < 1$ tel que pour tout $k \geq 1$, tout $y_{1:m}$ et toutes lois initiales ν et ν' de $(\mathsf{X}, \mathcal{X})$:*

$$\left\| \int_{\mathsf{X}} P(X_k \in . \mid X_1 = x, Y_{1:m} = y_{1:m}) [\nu(dx) - \nu'(dx)] \right\|_{VT} \leq 2\rho^k.$$

On reprend la preuve et les notations de la section 4.3 de [21], plus particulièrement du lemme 4.3.30, noté ici lemme 7.3.10. On va remontrer les deux premiers points du lemme avec l'hypothèse alternative, le reste de la preuve ne variant pas.

Notation 7.3.9. *On utilise dans le lemme les fonctions* backward *définies pour tous $k < m$, $x_k \in \mathsf{X}$, $A \in \mathcal{X}$ et $y_{k+1:m} \in \mathsf{Y}^{m-k}$ par :*

$$\beta_{k|m}[y_{k+1:m}](x_k) = \int_{x_{k+1}} \dots \int_{x_m} \prod_{l=k+1}^{m} Q(x_{l-1}, dx_l) g(x_l, y_l).$$

Lemme 7.3.10. *Sous l'hypothèse 7.3.7, les points suivants sont vérifiés.*

1. *Pour tous $1 \leq k \leq m$ et $x \in \mathsf{X}$,*
$$\varepsilon (g^-)^{m-k} \leq \beta_{k|m}[y_{k+1:m}](x) \leq (g^+)^{m-k}.$$
2. *Pour tout entier $0 \leq u < \lfloor m/d_0 \rfloor$ et pour toutes mesures de probabilités ν et ν' sur $(\mathsf{X}, \mathcal{X})$:*
$$\frac{\varepsilon}{\sigma^+} \left(\frac{g^-}{g^+}\right)^{d_0} \leq \frac{\int_{\mathsf{X}} d\nu(x) \beta_{ud_0|m}[y_{ud_0+1:m}](x)}{\int_{\mathsf{X}} d\nu'(x) \beta_{ud_0|m}[y_{ud_0+1:m}](x)} \leq \frac{\sigma^+}{\varepsilon} \left(\frac{g^+}{g^-}\right)^{d_0}.$$

Démonstration. Premier point. Soit $1 \leq k \leq m$ et $x \in \mathsf{X}$. Pour la borne inférieure, on écrit :

$$\begin{aligned}
\beta_{k|m}[y_{k+1:m}](x_k) &= \int_{x_{k+1}} \dots \int_{x_m} \prod_{l=k+1}^{m} Q(x_{l-1}, dx_l) g(x_l, y_l) \\
&\geq \left(\prod_{l=k+1}^{m} g^- \right) \int_{x_{k+1}} \dots \int_{x_m} \prod_{l=k+1}^{m} Q(x_{l-1}, dx_l) \mathbf{1}_{x_l \dashrightarrow y_l} \\
&\geq \varepsilon \prod_{l=k+1}^{m} (g^-)^{m-k}.
\end{aligned}$$

On a utilisé le deuxième et le troisième point de l'hypothèse pour respectivement l'avant-dernière et la dernière inégalité. La borne supérieure se traite directement en majorant $g(x_l, y_l)$ par g^+ pour tout l (avec le deuxième point de l'hypothèse).

Deuxième point. Soit $0 \leq u < \lfloor m/d_0 \rfloor$ et ν et ν' mesures de probabilités sur $(\mathsf{X}, \mathcal{X})$. Pour la borne inférieure, on écrit (en écrivant $\beta_{k|m}$ pour $\beta_{k|m}[y_{k+1:m}]$) :

$$\beta_{ud_0|m}(x_{ud_0}) = \int_{x_{ud_0+1:x_{(u+1)d_0}}} \prod_{i=ud_0+1}^{(u+1)d_0} Q(x_{i-1}, dx_i) g(x_i, y_i) \beta_{(u+1)d_0|m}(x_{(u+1)d_0}).$$

Par le deuxième point de l'hypothèse, on obtient :

$$\beta_{ud_0|m}(x_{ud_0}) \geq \left(\prod_{i=ud_0+1}^{(u+1)d_0} g^- \right) \int_{x_{(u+1)d_0}} D[y_{ud_0+1:(u+1)d_0}](x_{ud_0}, dx_{(u+1)d_0}) \beta_{(u+1)d_0|m}(x_{(u+1)d_0}).$$

Par le troisième point de l'hypothèse :

$$\beta_{ud_0|m}(x_{ud_0}) \geq \left(g^-\right)^{d_0} \varepsilon \int_{x_{(u+1)d_0}} d\lambda(x_{(u+1)d_0}) \beta_{(u+1)d_0|m}(x_{(u+1)d_0}).$$

On traite de façon similaire la borne supérieure et on obtient par quotient le résultat souhaité.

□

Réécriture du théorème avec l'hypothèse alternative.

Théorème 7.3.11. *On suppose que les hypothèses 7.1.4, 7.3.7, 7.1.6, 7.1.7 et 7.1.8 sont vérifiées et que de plus θ_0 est identifiable. Alors les points du théorème 7.1.9 sont vérifiés.*

Démonstration. On suit la preuve du chapitre 12 de [21] jusqu'à la section 12.5.5, avec la quantité $\rho = \max\left(1 - (g^-/g^+)^{d_0}\,\varepsilon/\sigma^+, 1-\varepsilon\right)$. □

Chapitre 8

Méthodes de simulation

Ce chapitre est dédié aux méthodes de simulations développées dans cette thèse. Il concerne principalement l'utilisation d'algorithmes particulaires pour calculer approximativement la log-vraisemblance d'observations provenant d'un modèle RN95+YpR. Ces approximations sont intéressantes car elles sont convergentes lorsque le nombre de particules utilisées tend vers l'infini (et de type Monte-Carlo), par opposition aux méthodes sans simulations comme les vraisemblances composites par approximations markoviennes.

Le cadre dans lequel on se place est celui des méthodes particulaires pour les chaînes de Markov cachées. Pour cela, après avoir fixé un modèle θ, on souhaite utiliser une structure séquentielle pour calculer récursivement la log-vraisemblance $\log p(y_{1:m})$ d'observations $y_{1:m}$ comme :

$$\log p(y_{1:m}) = \sum_{i=1}^{m} \log p(y_i|y_{1:i-1}).$$

L'utilisation d'une chaîne de Markov cachée provient du fait que la structure séquentielle n'est pas directement vérifiée par la suite d'observations mais par une suite de variables cachées $x_{1:m}$.

Les méthodes particulaires cherchent à simuler pour tout i et de façon récursive un échantillon selon la loi jointe de lissage $p(x_{1:i}|y_{1:i})dx_{1:i}$ dans le but d'évaluer une fonction bornée par rapport à cette loi. En choisissant convenablement cette fonction, on déduit une approximation de type Monte-Carlo de la quantité $p(y_i|y_{1:i-1})$ pour tout i, vérifiant des propriétés asymptotiques de convergence et de normalité. Ensuite, des théorèmes permettent d'établir la convergence et la normalité asymptotique de la log-vraisemblance recherchée $\log p(y_{1:m})$.

Pour mettre en œuvre des méthodes particulaires pour approcher la log-vraisemblance pour la classe de modèles RN95+YpR étudiée, on utilise la structure de chaîne de Markov cachée spécifique et les formes explicites des évolutions conditionnellement aux feuilles présentées dans le chapitre 6. En particulier, l'espace d'état des variables cachées correspond à l'ensemble des évolutions des dinucléotides encodés le long de l'arbre.

En dehors des algorithmes particulaires, on propose également une méthode de simulation exacte de la loi stationnaire.

Ce chapitre est découpé en trois parties. Dans la section 8.1, on effectue un rappel tel qu'exposé dans [21] des différents algorithmes particulaires dont l'algorithme SISR standard et le filtre particulaire auxiliaire ainsi que des propriétés de convergence associées. On décrit en particulier différents phénomènes de dégénérescence qui apparaissent lorsque le nombre de sites considérés devient important. Dans la section 8.2, on vérifie que l'on peut utiliser un filtre particulaire auxiliaire pour approcher de façon consistante et asymptotiquement normal la log-vraisemblance. Enfin dans la section 8.3, on propose un algorithme de simulation exacte de la loi stationnaire de type couplage par le passé.

8.1 Méthodes particulaires générales

On reprend les notations de la section 7.1. On choisit $(\mathsf{X}, \mathcal{X})$ et $(\mathsf{Y}, \mathcal{Y})$ deux espaces mesurables, ν une mesure de probabilité sur $(\mathsf{X}, \mathcal{X})$, Q un noyau de transition de $(\mathsf{X}, \mathcal{X})$ et G un noyau de transition de $(\mathsf{X}, \mathcal{X})$ vers $(\mathsf{Y}, \mathcal{Y})$ vérifiant pour une fonction g et une mesure de probabilité μ, pour tous $x \in X$ et $A \in \mathcal{Y}$:

$$G(x, A) = \int_A g(x, y)\mu(dy).$$

On considère la chaîne de Markov cachée $(X_i, Y_i)_{i\in 1:m}$ associée aux noyaux Q et G et à la loi initiale ν. On reprend les définitions suivantes de la section 3.1 de [21], pour tout $i \in [\![1, m]\!]$.

Définition 8.1.1. *La vraisemblance $L_i(y_{1:i})$ des observations $y_{1:i}$ s'écrit :*

$$\int_{x_1,\dots x_i} \nu(dx_1)g(x_1, y_1) \prod_{i=2}^{i} Q(x_{i-1}, dx_i)g(x_i, y_i).$$

Définition 8.1.2. *La loi de lissage jointe $\phi_{1:i|i}$ est définie, pour des observations $y_{1:i}$ et des états cachés $x_{1:i}$, par :*

$$\phi_{1:i|i}(y_{1:i}, dx_{1:i}) = L_i(y_{1:i})^{-1}\, \nu(dx_1)g(x_1, y_1) \prod_{i=2}^{i} Q(x_{i-1}, dx_i)g(x_i, y_i).$$

Définition 8.1.3. *La loi de filtrage ϕ_i est définie comme la loi conditionnelle de X_i sachant $Y_{1:i}$.*

Notation 8.1.4. *Pour simplifier la lecture, on note pour des observations $y_{1:m}$ et des états cachés $x_{1:m}$:*

- $p(x_{1:i}|y_{1:i})dx_{1:i} = \phi_{1:i|i}(y_{1:i}, dx_{1:i})$,
- $p(x_i|y_{1:i})dx_i = \phi_i(y_{1:i}, dx_i)$,
- $p(x_i|x_{i-1})dx_i = Q(x_{i-1}, dx_i)$,
- $p(x_i|x_{i-1}, y_i)dx_i = Q(x_{i-1}, dx_i)g(x_i, y_i)/\int_{x'} Q(x_{i-1}, dx')g(x', y_i)$.

Pour des observations $y_{1:m}$ fixées et des fonctions bornées $f_i : \mathcal{X}^i \to \mathbb{R}$ $(i \in [\![1, m]\!])$, le but de cette section est d'estimer pour $i \in [\![1, m]\!]$:

$$\bar{f}_i = \int f_i(x_{1:i})p(x_{1:i}|y_{1:i})dx_{1:i}.$$

8.1.1 Algorithme SISR générique

Pour cela, on suit la description de [65] qui introduit un algorithme d'échantillonnage séquentiel par importance avec rééchantillonnage générique (noté par la suite SISR générique). Les algorithmes utilisant l'échantillonnage par importance pour simuler contre une loi de filtrage ont été introduit en 1969 et 1970 dans [57, 58]. La phase de rééchantillonnage est introduite en 1993 dans [53].

On introduit une famille de densités de probabilités $(\varpi_i(x_{1:i}))_{i\in[\![1,m]\!]}$ et on cherche d'abord à estimer les fonctions f_i par rapport à $\varpi_i(x_{1:i})dx_{1:i}$. On introduit pour cela R_1 une mesure de probabilité et $(R_i)_{i\in 2:m}$ une famille de noyaux de transitions sur $(\mathsf{X}, \mathcal{X})$.

Algorithme 8.1.5. *SISR générique. On va simuler un échantillon de la loi $\varpi_i(x_{1:i})$ récursivement en i.*

L'algorithme dépend des paramètres suivants :
- *$(\varpi_i(x_{1:i}))_{i\in[\![1,m]\!]}$ densités de probabilités,*
- *$(R_i)_{i\in[\![1,m]\!]}$ mesure et noyaux instrumentaux,*
- *$l \in \mathbb{N}$ nombre de pas avant rééchantillonnage (aucun rééchantillonnage si $l = 0$),*
- *n nombre de particules.*

L'algorithme est décrit de la façon suivante : pour $i = 1$, pour $j \in [\![1, n]\!]$,
- *Simuler $x_1^{(j)}$ selon $R_1(.)$ (échantillonnage).*
- *Calculer $w_1(x_1^{(j)}) = \frac{\varpi_1(x_1^{(j)})}{R_1(x_1^{(j)})}$ et $W_1^{(j)} = \frac{w_1(x_1^{(j)})}{\sum_{k=1}^n w_1(x_1^{(k)})}$.*
- *Si $i \bmod l = 0$, simuler $\check{x}_1^{(j)}$ selon $\sum_{k=1}^n W_1^{(k)} \delta_{x_1^{(k)}}(dx_1)$ (rééchantillonnage) et poser $\hat{w}_1^{(j)} = 1$.*
- *Sinon, poser $\check{x}_1^{(j)} = x_1^{(j)}$ et $\hat{w}_1^{(j)} = w_1(x_1^{(j)})$.*

Pour $i \in [\![2, m]\!]$, pour $j \in [\![1, n]\!]$,
- *Simuler $x_i^{(j)}$ selon $R_i(. \mid \check{x}_{i-1}^{(j)})$ (échantillonnage).*
- *Calculer $w_i(\check{x}_{1:i-1}^{(j)}, x_i^{(j)}) = \hat{w}_{i-1}^{(j)} \frac{\varpi_i(\check{x}_{1:i-1}^{(j)}, x_i^{(j)})}{\varpi_{i-1}(\check{x}_{1:i-1}^{(j)}) R_i(x_i^{(j)} \mid \check{x}_{i-1}^{(j)})}$ et $W_i^{(j)} = \frac{w_i(\check{x}_{1:i-1}^{(j)}, x_i^{(j)})}{\sum_{k=1}^n w_i(\check{x}_{1:i-1}^{(k)}, x_i^{(k)})}$.*
- *Si $i \bmod l = 0$, simuler $\check{x}_{1:i}^{(j)}$ selon $\sum_{k=1}^n W_i^{(k)} \delta_{\check{x}_{1:i-1}^{(k)}, x_i^{(k)}}(dx_{1:i})$ (rééchantillonnage) et poser $\hat{w}_i^{(j)} = 1$.*
- *Sinon, poser $\check{x}_{1:i}^{(j)} = \left(\check{x}_{1:i-1}^{(j)}, x_i^{(j)}\right)$ et $\hat{w}_i^{(j)} = w_i(\check{x}_{1:i-1}^{(j)}, x_i^{(j)})$.*

Pour tout $i \in [\![1, m]\!]$, on approche enfin $\varpi_{i-1}(x_{1:i-1}) R_i(x_i \mid x_{i-1}) dx_{1:i}$ par :

$$\hat{\rho}_i^n(dx_{1:i}) = \frac{1}{n} \sum_{j=1}^n \delta_{\check{x}_{1:i-1}^{(j)}, x_i^{(j)}}(dx_{1:i})$$

et $\varpi_i(x_{1:i})dx_{1:i}$ par :

$$\hat{\varpi}_i^n(dx_{1:i}) = \sum_{j=1}^n W_i^{(j)} \delta_{\check{x}_{1:i-1}^{(j)}, x_i^{(j)}}(dx_{1:i}).$$

Remarque 8.1.6.

- *Si $l = 0$, on a pour $i \in [\![1, m]\!]$ que $i \bmod l \neq 0$ et on ne rééchantillonne jamais. Si $l = 1$, on rééchantillonne à chaque étape.*
- *La phase d'échantillonnage est quelquefois appelée phase de mutation et la phase de rééchantillonnage phase de sélection.*

8.1.2 Algorithme SISR standard

L'algorithme SISR standard correspond à choisir $\varpi_i(x_{1:i}) = p(x_{1:i}|y_{1:i})$. Deux choix de noyaux instrumentaux $(R_i)_{i\in 2:m}$ sont courants :

- Noyau prior : $R_i(dx_i|x_{i-1}) = p(x_i|x_{i-1})dx_i$, et $R_1 = \nu = p(x_1)$.
- Noyau optimal : $R_i(dx_i|x_{i-1}) = p(x_i|x_{i-1}, y_i)dx_i$ et $R_1 = p(x_1|y_1)$.

Dans les deux cas, $\bar{f}_i$ peut être estimé par :

$$\hat{f}^n_{i,\mathrm{SISR}} = \int f_i(x_{1:i})\hat{\varpi}^n_i(dx_{1:i}) = \sum_{j=1}^{n} W_i^{(j)} f_i(\check{x}^{(j)}_{1:i-1}, x_i^{(j)}).$$

8.1.3 Filtre particulaire auxiliaire

L'algorithme par filtre particulaire auxiliaire (noté par la suite APF) correspond à choisir $\varpi_i(x_{1:i}) = p(x_{1:i} \mid y_{1:i+1}) \propto p(x_{1:i} \mid y_{1:i})p(y_{i+1} \mid x_i)$. Cet algorithme a été introduit en 1999 dans [94]. Le choix le plus courant pour R_1 et les noyaux instrumentaux $(R_i)_{i\in 2:m}$ sont alors les noyaux optimaux. Dans ce cas, le nom d'échantillonnage i.i.d. est aussi utilisé (voir section 8.1.1 de [21]).

Contrairement à l'algorithme SISR standard, ce choix des ϖ_i ne permet pas d'obtenir un échantillon selon $p(x_{1:i} \mid y_{1:i})dx_{1:i}$ directement. On utilise alors la loi d'échantillonnage par importance $\hat{\rho}^n_i(dx_{1:i})$ pour approcher $\varpi_{i-1}(x_{1:i-1})R_i(x_i \mid x_{i-1})dx_{1:i}$. Il faut ensuite pondérer l'échantillon pour simuler selon $p(x_{1:i} \mid y_{1:i})$ depuis $\varpi_{i-1}(x_{1:i-1})R_i(x_i \mid x_{i-1})$. Les poids donnés pour $i \in [\![1, m]\!]$ et $j \in [\![1, n]\!]$ par :

$$\tilde{w}_i(\check{x}^{(j)}_{1:i-1}, x_i^{(j)}) = \frac{p(\check{x}^{(j)}_{1:i-1}, x_i^{(j)}|y_{1:i})}{\varpi_{i-1}(\check{x}^{(j)}_{1:i-1})R_i(x_i^{(j)}|\check{x}^{(j)}_{i-1})} \text{ et } \tilde{W}_i^{(j)} = \frac{\tilde{w}_i(\check{x}^{(j)}_{1:i-1}, x_i^{(j)})}{\sum_{k=1}^n \tilde{w}_i(\check{x}^{(k)}_{1:i-1}, x_i^{(k)})}.$$

$\bar{f}_i$ peut ainsi être estimée par :

$$\hat{f}^n_{i,\mathrm{APF}} = \sum_{j=1}^{n} \tilde{W}_i^{(j)} f_i(\check{x}^{(j)}_{1:i-1}, x_i^{(j)}).$$

Calcul de $\tilde{W}_i^{(j)}$ pour les noyaux optimaux. Dans le cas des noyaux optimaux pour R_i, on a :

$$\begin{aligned}\tilde{w}_i(x_{1:i}) &:= \frac{p(x_{1:i} \mid y_{1:i})}{\varpi_{i-1}(x_{1:i-1})R_i(x_i \mid x_{i-1})} \\ &= \frac{p(x_{1:i} \mid y_{1:i})}{p(x_{1:i-1} \mid y_{1:i})p(x_i \mid x_{i-1}, y_i)} \\ &= \frac{p(x_{1:i}, y_{1:i})}{p(x_{1:i-1}, y_{1:i})p(x_i \mid x_{i-1}, y_i)} \\ &= \frac{p(x_i \mid x_{1:i-1}, y_{1:i})}{p(x_i \mid x_{i-1}, y_i)} \\ &= 1.\end{aligned}$$

Ainsi, l'estimation de $\bar{f}_i$ précédente se simplifie en :

$$\hat{f}_{i,\mathrm{APF}}^n = \int f_i(x_{1:i})\hat{\rho}_i^n(dx_{1:i}) = \frac{1}{n}\sum_{j=1}^n f_i(\check{x}_{1:i-1}^{(j)}, x_i^{(j)}).$$

Remarque 8.1.7. *Remarque sur les noyaux prior avec le filtre particulaire auxiliaire. On calcule ici les poids $\tilde{w}_i^{(j)}$ dans le cas où les noyaux R_i utilisés sont les noyaux prior. On a :*

$$\tilde{w}_i(x_{1:i}) = \frac{p(x_{1:i}|y_{1:i})}{p(x_{1:i-1}|y_{1:i})p(x_i|x_{i-1})} = \frac{p(x_i|y_i, x_{i-1})}{p(x_i|x_{i-1})}.$$

Ainsi, on doit repondérer de façon à passer du noyau prior au noyau optimal. Il est donc préférable de considérer directement le noyau optimal pour le filtre particulaire auxiliaire plutôt que le noyau prior.

8.1.4 Résumé et résultats de convergence et de normalité asymptotique

Tableau récapitulatif des poids pour les méthodes particulaires courantes utilisées. Le tableau suivant résume pour chaque méthode, la loi cible utilisée ($\varpi_i(x_{1:i})$), les échantillons que l'on doit savoir simuler ($R_i(dx_i|x_{i-1})$) et les poids que l'on doit savoir calculer ($w_i(x_{1:i})$) pour mettre en œuvre la méthode.

	$\varpi_i(x_{1:i})$	$R_i(dx_i\|x_{i-1})$	$w_i(x_{1:i})$
SISR prior	$p(x_{1:i}\|y_{1:i})$	$p(dx_i\|x_{i-1})$	$p(y_i\|x_i)$
SISR optimal	$p(x_{1:i}\|y_{1:i})$	$p(dx_i\|x_{i-1}, y_i)$	$p(y_i\|x_{i-1})$
APF optimal	$p(x_{1:i}\|y_{1:i+1})$	$p(dx_i\|x_{i-1}, y_i)$	$p(y_{i+1}\|x_i)$

On remarque que dans les trois cas, la fonction de poids est bornée par 1.

Convergence et normalité asymptotique.

Théorème 8.1.8. *On fixe un site $i \in [\![1, m]\!]$. Pour $\hat{f}_i^n$ un des trois estimateurs de $\bar{f}_i$ parmi $\hat{f}_{i,SISR\ prior}^n$, $\hat{f}_{i,SISR\ optimal}^n$ et $\hat{f}_{i,APF\ optimal}^n$, la quantité suivante converge en loi vers une loi normale centrée de variance $\sigma^2(f_i)$:*

$$\sqrt{n}\left(\hat{f}_i^n - \bar{f}_i\right).$$

On note $\alpha_k(f_i) = \int f_i(x_{1:i})p(x_{k+1:i}|y_{k+1:i}, x_k)dx_{k+1:i} - \bar{f}_i$ *pour* $k \in [\![1, i-1]\!]$ *et* $\alpha_i(f_i) = f_i(x_{1:i}) - \bar{f}_i$.

Les variances $\sigma^2(f_i)$ *associées aux trois estimateurs précédents sont données par :*

$$\sigma^2_{SISR\ prior}(f_i) = \sum_{k=1}^{i} \int \frac{p(x_{1:k}|y_{1:i})^2}{p(x_{1:k}|y_{1:k-1})} \alpha_k(f_i)^2 dx_{1:k},$$

$$\sigma^2_{SISR\ optimal}(f_i) = \sum_{k=1}^{i} \int \frac{p(x_{1:k}|y_{1:i})^2}{p(x_{1:k-1}|y_{1:k-1})p(x_k|y_k, x_{k-1})}, \alpha_k(f_i)^2 dx_{1:k}$$

$$\sigma^2_{APF\ optimal}(f_i) = \sum_{k=1}^{i} \int \frac{p(x_{1:k}|y_{1:i})^2}{p(x_{1:k}|y_{1:k})} \alpha_k(f_i)^2 dx_{1:k}. \tag{8.1}$$

Démonstration. On remarque que pour les trois estimateurs, la fonction de poids est bornée. Ainsi, les conditions de régularité du théorème 1 de [26] sont vérifiées et permettent d'établir la première partie du théorème. Le calcul de la variance est donnée dans la section 2.3 de [65]. □

Remarque 8.1.9.

- *En général, les résultats obtenus avec la méthode APF sont meilleurs qu'avec les méthodes SISR. Néanmoins, il existe des constructions où la méthode APF est moins efficace que la méthode SISR, voir le contre-exemple de la section 3 de [65].*
- *Une condition suffisante pour avoir* $\sigma^2_{APF\ optimal}(f_i) \leq \sigma^2_{SISR\ prior}(f_i)$ *est de vérifier l'inégalité suivante pour tout* $k \in [\![1, i]\!]$ *et pour tous les états* $x_{1:k}$ *vérifiant* $p(x_{1:k}|y_{1:i}) > 0$ *:*

$$p(y_k|x_k) \geq p(y_k|y_{1:k-1}).$$

8.1.5 Problèmes de dégénérescence

Dégénérescence des poids dans le cas sans rééchantillonnage. L'algorithme 8.1.5 générique comporte un paramètre de rééchantillonnage $l \in \mathbb{N}$. Il correspond au nombre de pas avant d'effectuer un rééchantillonnage selon les poids associés à chaque particule. Initialement, cette phase de rééchantillonnage n'existait pas (elle a été introduite dans [53]). Dans ce cas, en pratique, un problème de dégénérescence des poids apparaît lorsque le nombre de sites grandit, c'est-à-dire que la plupart des quantités $W_i^{(j)}$ vont être beaucoup plus petites que $1/n$ et les estimateurs $\hat{\rho}_i^n(dx_{1:i})$ et $\hat{\varpi}_i^n(dx_{1:i})$ vont mal approcher les distributions souhaitées.

Dans l'exemple 7.3.1 de [21], le cas la chaîne de Markov est i.i.d. est traité et conduit à une croissance exponentielle en i de la variance des estimateurs de $\bar{f}_i$.

Il est donc nécessaire d'effectuer un rééchantillonnage après un certain nombre de pas l, qui dépend de la chaîne considérée. On peut pour cela utiliser un critère mesurant le nombre de particules qui restent associées à un poids significatif au site i par exemple avec (issu de [77]) :

$$N_{\text{eff}}(i) = \left[\sum_{j=1}^{n} \left(W_i^{(j)} \right)^2 \right]^{-1}.$$

Dégénérescence de $\check{x}_1^{(j)}$ dans le cas avec rééchantillonnage. L'algorithme 8.1.5 générique avec rééchantillonnage permet de rééchantillonner les poids pour éviter le problème de dégénérescence des poids décrit dans le paragraphe précédent, ce qui permet ainsi l'utilisation des estimateurs de $\bar{f}_i$ pour des valeurs de i plus élevées.

Par contre, si on cherche maintenant à simuler un échantillon $\hat{x}_1^j$ selon $p(x_1|y_{1:i})dx_{1:i}$, l'algorithme précédent avec rééchantillonnage n'est pas convenable lorsque le nombre de sites augmente puisque l'échantillon créé n'est pas constitué de tirages indépendants et de plus, le nombre de termes identiques dans l'échantillon va croître en fonction de i. En effet, à chaque étape de rééchantillonnage $\check{x}_{1:i}^{(j)}$ est issue de la loi $\sum_{k=1}^n W_i^{(k)} \delta_{\check{x}_{1:i-1}^{(k)}, x_i^{(k)}}(dx_{1:i})$. On obtient alors le phénomène de coalescence des trajectoires, décrit dans [53] et représenté graphiquement dans la section 2.2 de [4].

Quand i tend vers l'infini, l'échantillon est constitué de n termes identiques. L'algorithme avec rééchantillonnage n'est donc pas utilisable pour simuler un échantillon selon $p(x_1|y_{1:i})dx_{1:i}$, et plus généralement selon $p(x_{1:i}|y_{1:i})dx_{1:i}$.

Remarque 8.1.10. *Lorsque i est petit, on utilise un algorithme sans rééchantillonnage pour simuler un échantillon de $p(x_1|y_{1:i})dx_{1:i}$, qui est alors constitué de termes indépendants. Ici petit peut être au sens du critère $N_{\mathit{eff}}(i)$. On verra dans la section 8.2.3 que la structure markovienne particulière associée au modèle RN95+YpR permet d'effectuer la simulation d'un échantillon de $p(x_{1:m}|y_{1:m})dx_{1:m}$ sous certaines conditions.*

Simulation sans dégénérescence. On propose ici une méthode pour effectuer les simulations sans dégénérescence, alternative aux méthodes en deux étapes proposées dans [32] et [33]. Cette méthode ne s'applique que dans un cas particulier restrictif, mais s'applique pour la structure markovienne du modèle RN95+YpR décrite dans la section 6.2. On pourra alors simuler un échantillon selon $p(x_{1:m}|y_{1:m})dx_{1:m}$ par les méthodes particulaires sans voir apparaître ce problème de dégénérescence.

On sait qu'avec la méthode particulaire sans rééchantillonnage, les poids sont multiplicatifs et la répartition des poids va dégénérer (quelques poids très forts et les autres très faibles) lorsque le nombre de sites augmente (sauf dans le cas où tous les poids sont identiques à partir d'un site).

On sait qu'avec la méthode particulaire avec rééchantillonnage, l'ancêtre va dégénérer (un seul ancêtre) lorsque le nombre de site augmente. Rééchantillonner non systématiquement ne corrige pas le problème.

Hypothèse et proposition utilisées. On note

$$K_{y_i} = \{(x_{i-1}, x_{i+1}); p(x_{i-1}, y_i, x_{i+1}) > 0)\}$$

et on considère l'hypothèse suivante :

Hypothèse 8.1.11. *Il existe y_i est tel que pour tous $(x_{i-1}, x_{i+1}) \in K_{y_i}$:*

$$p(x_{i+1}|x_{i-1}, y_i) = p(x_{i+1}) \quad \text{(H0)}.$$

De cette hypothèse on déduit certaines propriétés.

Propriété 8.1.12. *Si l'hypothèse 8.1.11 est vraie, on a sur K_{y_i} :*
- $p(x_{i+1}|y_{1:i}) = p(x_{i+1}|x_{i-1}, y_i)$ (H1).

– $p(y_{i+1:m}|y_{1:i}) = p(y_{i+1:m})$ (H2).
– $p(y_{1:i}|x_{i+1:m}) = p(y_{1:i})$ (H3).

Démonstration. Comme l'hypothèse 8.1.11 est vraie, on a sur K_{y_i} :

$$\frac{p(x_{i+1}, x_{i-1}, y_{1:i})}{p(x_{i-1}, y_{1:i})} = p(x_{i+1})$$

d'où en intégrant sur x_{i-1} :

$$\frac{p(x_{i+1}, y_{1:i})}{p(y_{1:i})} = p(x_{i+1})$$

ce qui donne d'une part $p(y_{1:i}|x_{i+1}) = p(y_{1:i})$ d'où (H3) et d'autre part $p(x_{i+1}|y_{1:i}) = p(x_{i+1})$ (*).

Avec (*) et (H0), on aboutit par suite à : (H1) $p(x_{i+1}|y_{1:i}) = p(x_{i+1}|x_{i-1}, y_i)$.

Enfin, on obtient (H2) avec :

$$\begin{aligned} p(y_{i+1:m}|y_{1:i}) &= \int_{x_{i+1}} p(y_{i+1:m}, x_{i+1}|y_{1:i}) dx_{i+1} \\ &= \int p(y_{i+1:m}|x_{i+1}, y_{1:i}) p(x_{i+1}|y_{1:i}) dx_{i+1} \\ &= \int p(y_{i+1:m}|x_{i+1}) p(x_{i+1}) dx_{i+1} \\ &= \int p(y_{i+1:m}, x_{i+1}) dx_{i+1} \\ &= p(y_{i+1:m}) \end{aligned}$$

où on a utilisé (*) pour la troisième égalité.

□

Proposition 8.1.13. *Si (H0) est vraie, on a sur K_{y_i} :*

$$p(x_{1:m}|y_{1:m}) = p(x_{1:i-1}|y_{1:i}) p(x_i|x_{i-1}, x_{i+1}, y_i) p(x_{i+1:m}|y_{i+1:m}).$$

Démonstration. On écrit $p(x_{1:m}|y_{1:m}) = p(x_{1:i}|y_{1:m}, x_{i+1:m}) p(x_{i+1:m}|y_{1:m})$. Or :

$$p(x_{1:i}|y_{1:m}, x_{i+1:m}) = p(x_{1:i}|y_{1:i}, x_{i+1})$$

donc :

$$p(x_{1:m}|y_{1:m}) = p(x_{1:i}|y_{1:i}, x_{i+1}) \frac{p(y_{1:i}|x_{i+1:m}) p(x_{i+1:m}, y_{i+1:m})}{p(y_{1:i}) p(y_{i+1:m}|y_{1:i})}.$$

Avec (H2) et (H3) on a $p(x_{1:m}|y_{1:m}) = p(x_{1:i}|y_{1:i}, x_{i+1}) p(x_{i+1:m}|y_{i+1:m})$. Or,

$$p(x_{1:i}|y_{1:i}, x_{i+1}) = p(x_{1:i-1}|y_{1:i}) p(x_i|x_{i-1}, x_{i+1}, y_i) \frac{p(x_{i+1}|x_{1:i-1}, y_{1:i})}{p(x_{i+1}|y_{1:i})}.$$

On en déduit la proposition avec (H1). □

Mise en œuvre. On note

$$S_{\text{coupe}} = \{i \in [\![1, m]\!]; y_i \text{ vérifie la propriété (H0)}\}$$

et

$$(S_{\text{coupe}}(k))_{k \in 1:K}$$

les éléments ordonnés de cet ensemble.

Hypothèse 8.1.14. *On suppose que l'écart maximal entre deux éléments consécutifs de S_{coupe} reste borné p.s. quand le nombre de sites m tend vers l'infini.*

Sous l'hypothèse 8.1.14, on utilise alors pour simuler un échantillon de $p(x_{1:m}|y_{1:m})dx_{1:m}$ l'algorithme suivant, qui ne dégénère pas si le nombre de particules est assez grand, même lorsque le nombre de sites tend vers l'infini.

Algorithme 8.1.15.

- *Pour $k \in [\![1, K]\!]$, appliquer un algorithme particulaire (sans ou avec rééchantillonnage) sur le morceau $[\![S_{coupe}(k) + 1, S_{coupe}(k + 1) - 1]\!]$ avec n particules. On note par x_i^j le terme simulé pour un site i et pour une particule j.*
- *Pour chaque particule $j \in [\![1, n]\!]$, pour $k \in [\![1, K]\!]$, simuler l'évolution au site $i = S_{coupe}(k)$ conditionnellement à l'observation en ce site et à (x_{i-1}^j, x_{i+1}^j).*

8.2 Méthodes particulaires pour la structure markovienne de RN95+YpR

On utilise la structure markovienne explicite décrite dans la section 6.2 et les notations utilisées dans la section 7.3. On dispose alors des espaces $(\mathsf{Z}, \mathcal{Z})$ et $(\mathsf{Z}(\mathsf{T}), \mathcal{Z}(T))$, du noyau de transition Q sur $(\mathsf{Z}, \mathcal{Z})$ et de la densité g associée au noyau de transition G de $(\mathsf{Z}, \mathcal{Z})$ vers $(\mathsf{Z}(\mathsf{T}), \mathcal{Z}(T))$ (définis dans la section 7.2). On considère la chaîne de Markov cachée $(Z_i, Z_i(T))_{i \in 1:m}$ associée aux noyaux Q et G et à la loi initiale stationnaire.

Pour des observations $z_{1:m-1}(T)$ fixées, on va utiliser les algorithmes de la section précédente pour des fonctions $f_i : \mathcal{X}^i \to \mathbb{R}$ particulières, nous permettant de calculer successivement les rapport de probabilités $p(z_{i+1}(T)|z_{1:i}(T))$ puis la vraisemblance $p(z_{1:m-1}(T))$.

8.2.1 Calcul des rapports de probabilités

Choix des fonctions f_i. Pour $i \in [\![1, m-1]\!]$, on choisit :

$$f_i : z_1, \ldots, z_i \mapsto p(z_{i+1}(T)|z_i).$$

f_i est bornée et $\bar{f}_i$ est donnée par :

$$\begin{aligned}
\bar{f}_i &= \int p(z_{i+1}(T)|z_i) p(z_{1:i}|z_{1:i}(T)) dz_{1:i} \\
&= \int p(z_{i+1}(T)|z_{1:i}, z_{1:i}(T)) p(z_{1:i}|z_{1:i}(T)) dz_{1:i} \\
&= \frac{1}{p(z_{1:i}(T))} \int p(z_{1:i}, z_{1:i+1}(T)) dz_{1:i} \\
&= p(z_{i+1}(T)|z_{1:i}(T)).
\end{aligned}$$

Ainsi, réussir à estimer les différents $\bar{f}_i$ permet d'estimer les rapports de probabilités $p(z_{i+1}(T)|z_{1:i}(T))$.

Algorithme utilisé. On va voir que l'on peut utiliser l'algorithme APF optimal décrit dans la section précédente, c'est-à-dire que l'on peut simuler les noyaux R_i et calculer les poids w_i.

Noyaux R_i. On a : $R_i(dz_i|z_{i-1}) = p(dz_i|z_{i-1}, z_i(T))$. D'après la section 6.2.5 et la proposition 6.2.15, on sait échantillonner selon ces noyaux.

Poids w_i. Pour l'algorithme APF optimal, on a : $w_i(z_{1:i}) = p(z_{i+1}(T)|z_i)$. D'après le corollaire 6.2.5, on sait calculer exactement ces poids.

Remarque 8.2.1. *Dans les deux cas, on a utilisé la connaissance du modèle à la racine. Pour un site $i \in [\![1, m]\!]$, on pose ι_i la loi du nucléotide à la racine au site i (sans aucun encodage). On a $\iota_i \in \sigma(z_{i-1}, z_i)$ et :*

$$\begin{aligned}P(z_i, \rho(\iota_i), \eta(\iota_{i+1}) \mid \rho(z_1), z_{2:i-1}, \rho(\iota_1), \iota_{2:i-1}, \eta(\iota_i))&\\ =P(z_i, \rho(\iota_i), \eta(\iota_{i+1}) \mid z_{i-1}, \rho(\iota_1), \iota_{2:i-1}\eta(\iota_i))&\\ =P(z_i \mid \rho(\iota_i), \eta(\iota_{i+1}), z_{i-1})P(\rho(\iota_i) \mid \rho(\iota_1), \iota_{2:i-1}\eta(\iota_i))P(\eta(\iota_{i+1}) \mid \rho(\iota_1), \iota_{2:i}).&\end{aligned}$$

Ainsi, en utilisant à la racine un modèle markovien ou la loi stationnaire comme modèle à la racine, on peut calculer à chaque pas de temps les probabilités des prochains dinucléotides encodés à la racine.

Remarque 8.2.2. *Pour notre chaîne de Markov cachée, on préfère utiliser l'algorithme APF optimal plutôt que l'algorithme SISR prior puisque ce dernier conduit à une variance supérieure dans tous les cas. En effet d'après la remarque 8.1.9, il suffit de voir que pour tout $k \leq i$, pour tous $z_{1:k}$ vérifiant $p(z_{1:k}|z_{1:i}(T)) > 0$, on a :*

$$1 = p(z_k(T)|z_k) \geq p(z_k(T)|z_{1:k-1}(T)).$$

Convergence et normalité asymptotique. On utilise ici le théorème 8.1.8 pour notre modèle particulier. On obtient :

Théorème 8.2.3. *On fixe un site $i \in [\![1, m]\!]$. Pour $\hat{f}_i^n = \hat{f}_{i,APF\ optimal}^n$, la quantité suivante converge en loi vers une loi normale centrée indépendante de n :*

$$\sqrt{n}\left(\hat{f}_i^n - p(z_{i+1}(T)|z_{1:i}(T))\right).$$

8.2.2 Calcul de la vraisemblance

Par la section précédente, on sait estimer les rapports de probabilités $p(z_{i+1}(T)|z_{1:i}(T))$, et ainsi, en calculant aussi $p(z_1(T))$, d'estimer par produit la vraisemblance $p(z_{1:m-1}(T))$. On pose alors :

Définition 8.2.4. *On effectue l'algorithme particulaire APF optimal avec n particules et un rééchantillonnage tous les $r \in \mathbb{N}$ pas (0 signifie que l'on ne rééchantillonne pas). L'estimateur associé de vraisemblance est noté : $\hat{L}_{n,r\text{-}partic}$.*

Par consistance de chacun des morceaux du produit, on sait que $\hat{L}_{n,r\text{-partic}}$ est un estimateur consistant de $p(z_{1:m-1}(T))$. Bien que les différents termes ne sont pas indépendants en général, l'estimateur $\hat{L}_{n,r\text{-partic}}$ est normal asymptotiquement (voir par exemple le corollaire 1 de [23]). On peut alors écrire le théorème suivant :

Théorème 8.2.5. *Pour un nombre de sites m fixé, $\hat{L}_{n,r\text{-partic}}$ est un estimateur consistant et asymptotiquement normal de $p(z_{1:m-1}(T))$ lorsque le nombre de particules n tend vers l'infini.*

Lorsque m n'est plus fixé initialement mais dépend linéairement du nombre de particules n, le théorème asymptotique 8.2.5 n'est plus valable et il est nécessaire de corriger l'estimateur par un biais proportionnel à m/n. On énonce la conjecture issue de [95] sur le comportement asymptotique associé à cet estimateur. Cette conjecture a été démontrée dans [13] pour des estimateurs particulaires associées à des chaînes de Markov vérifiant entre autres la condition de positivité stricte de la densité g associée aux transitions de la variable cachée à la variable des observations.

Pour la chaîne de Markov utilisée ici (voir section 8.2), cette condition de positivité n'est pas vérifiée. Néanmoins, la validité du théorème dans notre cadre semble extrêmement plausible.

Conjecture 8.2.6. *On suppose que le nombre de sites m évolue en fonction du nombre de particules n de la façon suivante :*

$$\lim_{m\to+\infty} \frac{m}{n} =: \alpha \in]0, +\infty[.$$

De plus, on suppose que la variance v_m de l'estimateur $\hat{L}_{n,r\text{-partic}}$ vérifie :

$$\lim_{m\to+\infty} \alpha^{-1} v_m =: \sigma^2 \in]0, +\infty[.$$

Alors l'estimateur de log-vraisemblance $\hat{L}_{n,r\text{-partic}}$ converge en loi vers une distribution normale d'espérance $-\frac{1}{2}\alpha\sigma^2$ et de variance $\alpha\sigma^2$.

8.2.3 Solution aux problèmes de dégénérescence

Dans le dernier paragraphe de la section 8.1.5, l'hypothèse supplémentaire 8.1.11 permet de simuler un échantillon des états cachés conditionnellement aux observations.

Pour la chaîne de Markov cachée associée à la structure markovienne explicite du modèle RN95+YpR, cette hypothèse est vérifiée pour tous les sites qui ont RY sur chacune des feuilles (voir le corollaire 3.5.2).

Si de plus l'hypothèse 8.1.14 est vérifiée, alors l'algorithme 8.1.15 peut être utilisé pour simuler un échantillon selon $p(z_{1:m}|z_{1:m}(T))dz_{1:m}$. Cette hypothèse est acceptable pour les deux séquences génomiques étudiées dans la section 10.6.

8.2.4 Calcul de la vraisemblance au maximum de vraisemblance

Pour des observations fixées et un modèle M inclus dans le modèle RN95+YpR, on souhaite calculer la vraisemblance au maximum de vraisemblance. On va considérer différents algorithmes, du moins efficace au plus efficace.

Tout d'abord, on choisit une méthode $\hat{L}$ pour le calcul de la vraisemblance. On peut choisir une méthode particulaire comme $\hat{L}_{n,r\text{-partic}}$, consistante quand le nombre de particules tend vers l'infini (voir définition 8.2.4 et théorème 8.2.5). On peut aussi choisir l'approximation markovienne à l pas $\hat{L}_{l\text{-Markov}}$ (voir définition 4.2.3), qui n'est pas consistante à l fixé.

On peut alors utiliser l'algorithme suivant :

Algorithme 8.2.7.

1. *Estimer le maximum de vraisemblance $\hat{\theta}_0$ de la séquence sous le modèle M par la méthode $\hat{L}$.*
2. *Calculer la vraisemblance associée à ce coefficient $\hat{\theta}_0$.*

Cet algorithme est difficilement utilisable pour $L = \hat{L}_{n,r\text{-partic}}$ et des séquences longues. En effet, pour obtenir une approximation crédible de la vraisemblance en un paramètre dans ce cas, il faut un nombre important de particules n, coûteux en temps de calcul.

On fait alors intervenir la méthode par triplets encodés, qui ne permet pas de calculer la vraisemblance du modèle mais d'estimer de façon consistante et asymptotiquement normale le maximum de vraisemblance du modèle. On obtient l'algorithme suivant

Algorithme 8.2.8.

1. *Estimer le maximum de vraisemblance $\hat{\theta}_0$ de la séquence sous le modèle* M *par la méthode des triplets encodés (voir section 4.1.1).*
2. *Calculer la vraisemblance associée à ce coefficient $\hat{\theta}_0$ par la méthode $\hat{L}$.*

Cet algorithme permet de réduire à un le nombre de calculs utilisant la méthode $\hat{L}$ et est utilisable en pratique avec $L = \hat{L}_{n,r\text{-partic}}$.

On peut améliorer l'algorithme 8.2.8 en se rappelant du corollaire 3.5.2 de découpage en morceaux indépendants : si pour un couple de sites, toutes les observations sont égales au dinucléotide encodé RY, on peut séparer en deux morceaux indépendants le calcul de la vraisemblance. On a alors l'algorithme suivant pour un entier $k \geq 2$ et une méthode d'approximation de la vraisemblance $\hat{L}$:

Algorithme 8.2.9.

1. *Estimer le maximum de vraisemblance $\hat{\theta}_0$ de la séquence sous le modèle* M *par la méthode des triplets encodés (voir section 4.1.1).*
2. *Découper en morceaux indépendants grâce au corollaire 3.5.2.*
3. *Pour chaque morceau :*
   - *Si le nombre de nucléotides est inférieur ou égal à k, calculer exactement la vraisemblance de ce morceau.*
   - *Sinon, calculer par la méthode $\hat{L}$ une approximation de la vraisemblance du morceau.*
4. *Faire le produit des vraisemblances de chaque morceau pour obtenir une approximation de la vraisemblance au maximum de vraisemblance.*

L'algorithme calcule ainsi une partie de la vraisemblance de manière exacte (pour tous les morceaux de longueur plus petite que k) et l'autre partie avec la méthode $\hat{L}$. En pratique, on choisit k entre 3 et 5.

Cet algorithme est plus efficace que l'algorithme 8.2.8 dans le cas où les séquences associées aux feuilles varient peu, ou bien dans le cas de petits arbres. En effet dans ce cas le nombre de morceaux est plus important. On utilisera cet algorithme pour la comparaison de modèles dans la section 10.6 sur deux jeux de données.

Exemple 8.2.10. *Pour donner une idée du gain d'efficacité dans le cas où les feuilles varient peu, on considère une séquence à valeurs dans $\{R, Y\}$, avec p la probabilité de valoir R et $q = 1 - p$ la probabilité de valoir Y. On remarque que cette séquence peut être vue comme la loi stationnaire π-encodée d'un modèle RN95+YpR vérifiant $p = (v_A + v_G)/(v_A + v_C + v_G + v_T)$.*

Partant d'un nucléotide π-encodé Y, la loi du nombre de sites à parcourir avant le motif RY est donnée, pour $k \geq 2$ par :

$$\begin{cases} \frac{pq}{p-q}(p^{k-1} - q^{k-1}) & si\ p \neq q, \\ (k-1)p^k & si\ p = q. \end{cases}$$

L'espérance est donnée par $1/pq$, et on peut alors calculer la proportion de nucléotides qui est calculée de manière exacte, suivant les valeurs choisies de p et de k. On regroupe dans le tableau 8.1 différentes valeurs de p et de k.

$p \setminus k$	3	4	5	6
0.1	4%	7%	10%	14%
0.2	13%	21%	30%	38%
0.3	22%	36%	49%	60%
0.4	29%	46%	61%	73%
0.5	31%	50%	66%	77%

TABLE 8.1 – Proportion de nucléotides calculé de manière exacte dans le cadre de l'exemple 8.2.10, en fonction du pas k et de p.

8.3 Simulation exacte de la loi stationnaire

On cherche à simuler exactement une séquence Φ-encodée de longueur m selon la loi stationnaire d'un modèle RN95+YpR. Pour cela, on va appliquer un algorithme de type couplage par le passé (voir l'algorithme de Propp et Wilson [98]) et utiliser des notations proches de celles issues de [14].

L'idée est de considérer d'abord l'évolution dans l'encodage π, qui d'après la proposition 3.2.7 est décrite explicitement par la matrice de taux de sauts de la définition 3.2.4 :

$$Q_\pi = \begin{array}{c} \\ R \\ Y \end{array} \begin{array}{c} \begin{array}{cc} R & Y \end{array} \\ \begin{pmatrix} . & v_Y \\ v_R & . \end{pmatrix} \end{array}$$

avec $v_Y = v_C + v_T$ et $v_R = v_A + v_G$.

On utilise ensuite la structure de la section 6.3 basée sur le π-encodage qui permet de désambigüiser de façon indépendante chacun des dinucléotides se chevauchant $(Z_i)_{i \in [\![1, m-1]\!]}$ (d'après la proposition 6.3.1).

8.3.1 Dynamique observée à travers les processus de Poisson

On décrit ici la dynamique π-encodée sans conditionnement sous forme d'une superposition de processus de Poisson.

Pour chaque site i de la séquence, on considère deux processus de Poisson homogènes sur $\mathbb{R}$, indépendants (entre eux et entre les différents sites), appelés $\mathcal{R}_i$ et $\mathcal{Y}_i$. Le taux des processus $\mathcal{R}_i$ est v_R et celui des processus $\mathcal{Y}_i$ est v_Y.

Pour chaque processus de Poisson, des sonneries retentissent de façon i.i.d. selon une loi exponentielle de paramètre associé au taux. Lorsque qu'une sonnerie $\mathcal{R}_i$ (resp. $\mathcal{Y}_i$) se produit à l'instant t, on substitue la lettre présente au site i à l'instant t par R (resp. par Y).

On a alors une description de la dynamique π-encodée sans conditionnement et on note ξ une réalisation de cet ensemble de sonneries. On se reporte à [14] pour des détails supplémentaires de ce type de construction par processus de Poisson.

8.3.2 Verrouillage des sites

On fixe ici l'ensemble des sonneries ξ. On cherche des instants où la dynamique est verrouillée, selon la définition suivante :

Définition 8.3.1. *On dit que le couple $(i, i+1)$ est verrouillé aux instants (t_1, t_2) (avec $t_1 < t_2$) pour l'ensemble ξ si pour tous instants $s \leq t_1$ et $t \geq t_2$, la dynamique de Z_i est identique à partir de l'instant t_2, partant de n'importe quelle condition initiale à l'instant s et soumise aux sonneries ξ.*

Proposition 8.3.2. *Si l'une des deux conditions suivantes sur ξ est vérifiée, alors le couple $(i, i+1)$ est verrouillé aux instants (t_1, t_2)*

- *une sonnerie $\mathcal{R}_i$ retentit en t_1, une sonnerie $\mathcal{Y}_{i+1}$ retentit en t_2 et aucune sonnerie n'a lieu sur $]t_1, t_2[$ pour les sites i et $i+1$.*
- *une sonnerie $\mathcal{Y}_{i+1}$ retentit en t_1, une sonnerie $\mathcal{R}_i$ retentit en t_2 et aucune sonnerie n'a lieu sur $]t_1, t_2[$ pour les sites i et $i+1$.*

Démonstration. Si l'une des deux conditions est vérifiée, alors partant de n'importe quelle condition initiale à l'instant s, le dinucléotide encodé dans π est RY à l'instant t_2. Or, après désambigüisation, le dinucléotide encodé dans (ρ, η) à cet instant est encore RY, qui est bien indépendant de la condition initiale à un instant $s < t_1$ choisie. □

La proposition suivante permet de connaître la loi du temps avant verrouillage d'un couple $(i, i+1)$.

Proposition 8.3.3. *On superpose les quatre processus de Poisson $\mathcal{R}_i$, $\mathcal{Y}_i$, $\mathcal{R}_{i+1}$ et $\mathcal{Y}_{i+1}$ et lors d'une sonnerie, les probabilités de sortie sont :*

- *une sonnerie $\mathcal{R}_i$ avec probabilité $\frac{v_R}{2(v_R+v_Y)}$ (mouvement A).*
- *une sonnerie $\mathcal{Y}_i \cup \mathcal{R}_{i+1}$ avec probabilité $\frac{1}{2}$ (mouvement B).*
- *une sonnerie $\mathcal{Y}_{i+1}$ avec probabilité $\frac{v_Y}{2(v_R+v_Y)}$ (mouvement C).*

On note $a = \frac{v_R}{2(v_R+v_Y)} \in]0, 1/2[$ puis on pose :

$$\nu = \left[\left(\frac{1}{2}+a\right)a, \frac{1}{2}, (1-a)\left(\frac{1}{2}-a\right), 2a\left(\frac{1}{2}-a\right)\right],$$

$$M = \begin{array}{c} \\ \{A,B\}A \\ \{A,B,C\}B \\ \{B,C\}C \\ AC \cup CA \end{array} \begin{array}{c} \begin{array}{cccc} \{A,B\}A & \{A,B,C\}B & \{B,C\}C & AC \cup CA \end{array} \\ \begin{pmatrix} a & 1/2 & 0 & 1/2-a \\ a & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2-a & a \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

Pour tout $j \geq 2$, on note $f(j)$ la quatrième composante du vecteur νM^{j-2}.

Le nombre de sonneries L avant verrouillage d'un couple $(i, i+1)$ vérifie pour tout $j \geq 2$:

$$P(L \leq j) = f(j).$$

De plus, en posant $\lambda = 2(v_Y + v_R)$, le temps d'attente T avant verrouillage vérifie pour tout $t \geq 0$:

$$P(T \leq t) = \sum_{j \geq 2} e^{-\lambda t} \frac{(\lambda t)^j}{j!} P(L \leq j).$$

Démonstration. Trouver le nombre de sonneries avant verrouillage correspond à trouver le nombre de lettres à écrire avant d'obtenir le motif AC ou CA lorsque les lettres sont tirés parmi A, B et C avec probabilités respectivement a, $1/2$ et $1/2 - a$. ν représente alors la loi d'un couple de lettres et M la dynamique avec les états AC et CA comme cimetières.

Cela donne la première partie de la proposition. La deuxième partie s'en déduit puisque le nombre de sonneries sur un intervalle $[0, t]$ suit une loi de Poisson de paramètre λt. □

8.3.3 Algorithme de simulation de la loi stationnaire

La proposition de verrouillage 8.3.2 et la proposition 6.3.1 permettent de construire un algorithme de simulation exacte de la loi stationnaire :

Algorithme 8.3.4. *Pour tout $i \in [\![1, m-1]\!]$:*

- *À partir du temps initial 0, on construit vers le passé les marqueurs de sonneries de la dynamique π-encodée des sites i (ceux qui n'existent pas encore) et $i+1$, jusqu'à ce que le couple $(i, i+1)$ soit verrouillé. On note alors les instants de verrouillage (t_1^i, t_2^i).*
- *On décrit l'évolution dans π à partir de la condition initiale $\pi_i(t_2^i) = RY$ et selon les marqueurs de sonneries créés, de l'instant t_2^i à l'instant 0.*
- *On désambigüise à l'aide de la proposition 6.3.1 et on obtient une description de l'évolution dans (ρ, η).*

À l'instant 0, on obtient alors une séquence de dinucléotides encodés, de laquelle on déduit une séquence Φ-encodée de longueur m.

Remarque 8.3.5. *Cet algorithme est différent des algorithmes de type couplage par le passé classiques puisqu'il n'impose pas de choisir un temps négatif initial à partir duquel on effectue l'évolution jusqu'au temps 0. Ici, on construit vers le passé les marqueurs de sonneries de la dynamique π-encodée et on s'arrête donc au premier instant où on sait que la coalescence a lieu.*

Chapitre 9

Implémentation

Ce chapitre décrit le programme `particulaire` permettant de calculer numériquement $\hat{L}_{n,r\text{-partic}}$ (voir définition 8.2.4), c'est-à-dire une approximation particulaire de la log-vraisemblance d'observations issues d'un modèle RN95+YpR.

Le chapitre est découpé en trois parties. Dans la section 9.1, on explique comment exécuter le programme du point de vue de l'utilisateur, sans chercher à comprendre comment fonctionne le programme. Dans la section 9.2, on décrit la structure générale du programme, en donnant en particulier un diagramme des différentes classes créées. Enfin, dans la section 9.3, on détaille les algorithmes importants utilisés pour concevoir le programme.

Remarque 9.0.6. *Dans ce chapitre, le mot* méthode *prend le sens de fonction et le mot* attribut *prend le sens de variable.*

9.1 Programme du point de vue de l'utilisateur

Les entrées du programme sont les suivantes :
- le jeu de séquences de nucléotides,
- l'arbre considéré (la topologie de l'arbre et les différentes longueurs de branches),
- les paramètres du modèle d'évolution,
- la loi à la racine (l'implémentation s'est limitée à un modèle markovien à un pas pour les dinucléotides encodés),
- r le nombre de pas avant rééchantillonnage (0 pour faire l'évolution sans rééchantillonnage),
- K le nombre de répétitions, i.e. le nombre d'évolutions particulaires indépendantes exécutées,
- n le nombre de particules utilisées pour chaque répétition.

Les sorties du programme sont :
- la log-vraisemblance approchée du jeu de séquences observé, obtenue pour chaque répétition,
- la log-vraisemblance approchée de $P(z_i(T)|z_{1:i-1}(T))$, obtenue pour chaque site et pour chaque répétition.

Les entrées correspondant aux jeux de séquences, à l'arbre, aux paramètres du modèle d'évolution et à la loi à la racine, ainsi que les sorties, sont sous la forme de fichiers textes. Globalement, l'exécution est réalisée par une commande de la forme :

```
./particulaire dossierSequences dossierEntrees dossierSorties nomSeq r K n.
```

9.2 Structure générale du programme

Le programme a été réalisé en C++. Un diagramme simplifié des différentes classes et de leurs relations est représenté sur la figure 9.1.

Le programme utilise deux classes externes :
- `Eigen` [55], une bibliothèque de templates C++ d'algèbre linéaire,
- `Mersenne Twister` [80], un générateur de nombres pseudo-aléatoires.

On note m la longueur des séquences observées. Un site $i \in [\![1, m-1]\!]$ reprend l'indexation des variables (Z_i) et est associé à l'évolution d'un dinucléotide encodé. À partir de la classe `main`, la structure générale est la suivante :
- la classe `main` utilise la classe d'entrées-sorties `IO` pour importer les fichiers externes (jeux de séquences, arbre, paramètres du modèles et loi à la racine). Ensuite, elle crée une instance de la classe `EvolParticulaire` et lui demande d'effectuer la méthode *évolution particulaire*. Enfin elle exporte les approximations de log-vraisemblances à l'aide de la classe `IO`.
- la classe `EvolParticulaire` est constituée de n instances de `EvolArbre` (une pour chaque particule). La méthode *évolution particulaire* consiste à effectuer, pour chaque site $i \in [\![1, m-1]\!]$,
  1. la méthode *évolution* de chaque instance de `EvolArbre`,
  2. le rééchantillonnage des particules (utilise `Alea`),
  3. la création pour le site $i+1$ des dinucléotides encodés associés à la racine de l'arbre (utilise `Alea` et `LoiRacine`).
- la classe `EvolArbre` est constituée à priori de m instances de `EvolSiteArbre` (une pour chaque site). Pour diminuer le coût en mémoire, on se restreint à deux instances, associées aux sites $i-1$ et i. La méthode *évolution* consiste à effectuer la méthode *évolution le long de l'arbre* de l'instance associée au site i de `EvolSiteArbre`, sachant l'évolution au site $i-1$.
- la classe `EvolSiteArbre` est constituée des instances `Arbre` et `EvolSite` pour chaque arête de l'arbre. La méthode *évolution le long de l'arbre* remplit les différentes matrices d'évolutions (non aléatoire) puis effectue l'évolution de toutes les instances `EvolSite` (utilise la classe `Alea`).
- la classe `EvolSite` est constituée d'une instance de `Noyau`, des instants et valeurs de changement dans l'alphabet des dinucléotides encodés pour le site en cours, ainsi que des instants et valeurs de changement dans l'alphabet $\{R, Y\}$ pour le site précédent. Les méthodes incluses dans cette classe correspondent aux méthodes ne nécessitant pas la connaissance complète de l'arbre.
- la classe `Noyau` est constituée d'une instance de `MatricesDeSaut`. Elle regroupe toutes les méthodes qui relient les matrices de sauts aux alphabets utilisés.

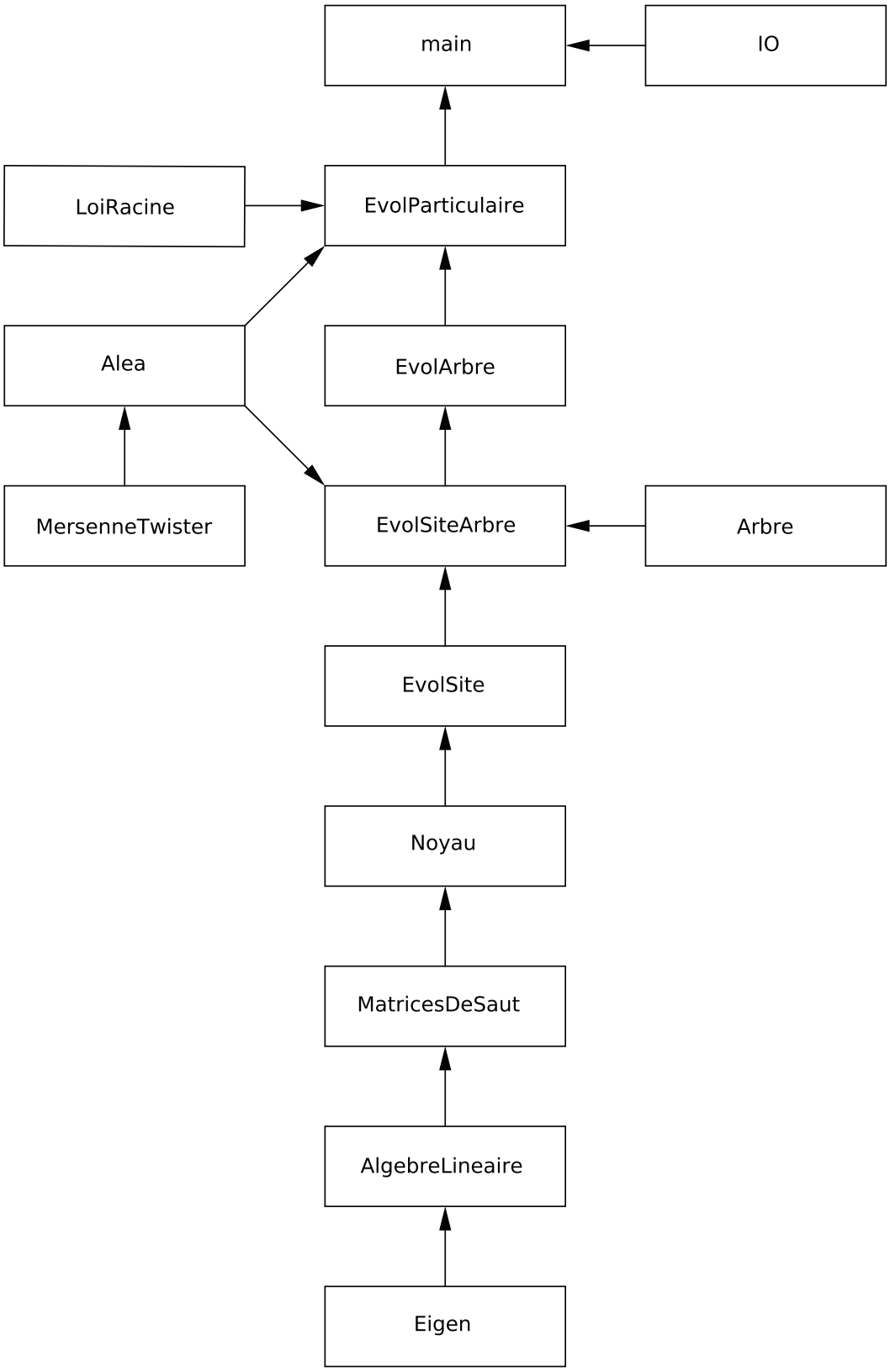


FIGURE 9.1 – Diagramme simplifié des différentes classes et de leurs relations pour le programme d'évolution particulaire.

– la classe `MatricesDeSaut` regroupe les matrices $W_Y, W_R, U_{Y\to R}$ et $U_{R\to Y}$ du théorème 6.2.3 (respectivement de taille 6×6, 3×3, 6×3 et 3×6).

9.3 Structure détaillée

On décrit dans cette section la structure détaillée du programme. Pour donner une idée du temps de calcul des méthodes décrites, on exécute `particulaire` avec les paramètres $r = 1$, $K = 10$ et $n = 1000$, puis on note le temps passé dans chacune des méthodes à l'aide des logiciels Valgrind [87] et KCachegrind [118]. Les pourcentages donnés dans cette section se réfèrent à cette exécution. Ces pourcentages sont exprimés par rapport à l'exécution de l'ensemble du programme (qui représente alors 100%), et correspondent à la proportion de temps passé dans la méthode et dans l'ensemble des méthodes appelées par cette méthode.

Notons également que les méthodes décrites dans les sections 9.3.1 et 9.3.2 ne font pas intervenir le générateur de nombre pseudo-aléatoire, contrairement à celles des sections suivantes 9.3.3 et 9.3.4.

9.3.1 Calculs matriciels

Exponentielles de matrices.

Pour calculer l'exponentielle d'une matrice Q, une procédure expérimentée a été de diagonaliser la matrice (sous réserve qu'elle soit effectivement diagonalisable) pour obtenir :

$$Q = PDP^{-1},$$

avec D matrice diagonale et P inversible, puis d'en déduire pour $t \in \mathbb{R}$:

$$e^{tQ} = Pe^{tD}P^{-1}. \tag{9.1}$$

Ainsi, en sauvegardant P^{-1}, P et D, on en déduit e^{tQ} (pour $t \in \mathbb{R}$) par produit de 3 matrices. Lorsque la matrice Q est proche d'une matrice singulière, des erreurs d'approximation apparaissent et peuvent être importantes (voir [83] paragraphe 6. *Matrix decomposition methods*).

Pour éviter ces problèmes potentiels, on remplace cette procédure par celle déjà implémentée dans la classe `Eigen` pour calculer l'exponentielle de matrice. Cette dernière utilise une méthode *scaling and squaring method* puis un approximant de Padé (méthode implémentée à partir de [60]). La complexité de cette méthode est de l'ordre de k^3, où $k \times k$ est la taille de la matrice considérée.

Pour accélérer l'exécution du programme, on choisit de fixer un intervalle de temps $\varepsilon > 0$ et de calculer e^{tQ} par pas de ε pour t allant de 0 jusqu'à la longueur de l'arête la plus longue de l'arbre. Dans les applications du chapitre 10, le pas choisi est $\varepsilon = 10^{-5}$.

Temps d'exécution : 1% pour `Noyau::`*remplirExponentiellesMatricesDeSaut*.

Produits le long d'une arête.

En supposant que l'évolution π-encodée au site i est connue, on exprime la matrice de transition de l'évolution Z_i sur une arête par le corollaire 6.2.5 (de l'instant initial de

l'arête jusqu'à l'instant final de l'arête). On enregistre les résultats obtenus pour toutes les arêtes.

Exemple 9.3.1.

t	π_i	Z_i
0	Y	
0.2	R	
0.3	Y	
0.9	R	
1	R	

Pour cet exemple, on calcule la matrice de transition (avec les notations du théorème 6.2.3) :

$$e^{\Delta_1 W_Y} U_{Y\to R} e^{\Delta_2 W_R} U_{R\to Y} e^{\Delta_3 W_Y} U_{Y\to R} e^{\Delta_4 W_R} = e^{0.2 W_Y} U_{Y\to R} e^{0.1 W_R} U_{R\to Y} e^{0.6 W_Y} U_{Y\to R} e^{0.1 W_R}.$$

Temps d'exécution : 26% pour la méthode `EvolSite::`*matriceDEvolution*, qui correspond à accéder aux exponentielles de matrices déjà calculées et à effectuer les produits matriciels.

Produits le long de l'arbre – algorithme de Felsenstein.

On suppose que les feuilles $z_i(T)$ pour le site i sont connues (correspondant à un ensemble de dinucléotides encodés). On reprend les notations décrites dans la section 6.4.2. Chaque arête est associée à son nœud fils v, et pour tout nœud v, on note par $l(v)$ l'ensemble des nœuds feuilles de l'arbre issus de v. De plus, on note $G_{v\to l(v)}$ la matrice de transition sur l'arbre à partir de l'arête v.

D'après l'algorithme de Felsenstein décrit dans la section 6.4.2, on déduit pour chaque arête v de l'arbre le vecteur défini par :

$$z \mapsto G_{v\to l(v)}(z, z_i(T)).$$

L'ensemble de ces vecteurs est calculé récursivement, en partant des arêtes associées aux feuilles puis en remontant dans l'arbre jusqu'à la racine. L'espace de départ de la fonction dépend de la valeur de π_i :

- si $\pi_i = Y$, alors $z \in \{CA, CG, CY, TA, TG, TY\}$,
- si $\pi_i = R$, alors $z \in \{RA, RG, RY\}$.

Temps d'exécution : 13% pour la méthode `EvolSiteArbre::`*remplirMatricesVecteurs*. L'attribut rempli est *m_ vecteursDEvolution*.

9.3.2 Calculs pour la fonction de survie

D'après la section 6.2.6, on rappelle que la fonction de survie à partir d'un instant t_0 s'exprime de la façon suivante :

Proposition 9.3.2. *Pour $t \in [t_0, t_1[$ (où t_1 est défini selon la définition 6.2.4) et z dinucléotide encodé, la valeur de la fonction de survie $\bar{F}_{t_0}(t)(z)$ est donnée par :*

$$\exp\left(-(t-t_0)\sum_{z'\neq z}\hat{Q}_{t_0,\pi_i}(z,z')\right)\frac{G(t)(z,z_T)}{G(t_0)(z,z_T)}.$$

Cette fonction est décroissante, vaut 1 en t_0 et vaut une valeur strictement positive p_1 en t_1-. Pour $t \geq t_1$, la fonction est nulle.

Pour implémenter la méthode `EvolSiteArbre::`*fonctionSurvieArbre* associée à cette fonction, on calcule d'une part le membre de gauche $\exp\left(-(t-t_0)\sum_{z'\neq z}\hat{Q}_{t_0,\pi_i}(z,z')\right)$ et d'autre part chaque terme du membre de droite $\frac{G(t)(z,z_T)}{G(t_0)(z,z_T)}$.

Calcul du membre gauche de la fonction.

Le membre de gauche se calcule sans avoir besoin de connaître l'ensemble de l'évolution, directement à partir de la formule.

Temps d'exécution : 4% pour la méthode `EvolSite::`*membreGaucheSurvie*.

Calcul des termes du membre droit de la fonction.

Le calcul effectif des termes de la forme $G(t)(z, z_T)$ fait intervenir l'ensemble de l'arbre (à partir de l'instant t), et se calcule à l'aide des attributs *m_ vecteursDEvolution*.

Temps d'exécution : 29% pour `EvolSiteArbre::`*membreDroitSurvieMorceau*, dont 26% pour la méthode appelée `EvolSite::`*matriceDEvolution*.

Pseudo-inverse de la fonction de survie.

Le calcul d'une pseudo-inverse $u \mapsto \bar{F}^{\leftarrow}(u)$ de la fonction de survie se fait en deux étapes :

- si $u \leq p_1$, alors u n'a pas d'antécédent par $\bar{F}$ et on pose $\bar{F}^{\leftarrow}(u) := -1$
- sinon, u admet un unique antécédent. On obtient un instant t_c approchant cet antécédent en procédent par dichotomie. La précision est fixée par la valeur précision $= 10^{-3}$, de telle sorte que :

$$u \in [\bar{F}(t_c) - \text{précision} \ ; \ \bar{F}(t_c) + \text{précision}].$$

La méthode correspondante est `EvolSiteArbre::`*invFonctionChangementArbre*.

9.3.3 Description d'une étape de l'avancement d'un site

On se place sur une arête de l'arbre à l'instant t_0. On suppose connu l'état z_0 du dinucléotide encodé en cet instant (parmi $\{CA, CG, CY, TA, TG, TY\}$ ou $\{RA, RG, RY\}$ suivant $\pi_i(t_0)$).

Pour effectuer une étape de l'avancement de l'arbre, on simule u provenant d'une loi uniforme sur $]0, 1[$. On effectue ensuite les étapes suivantes :

- à l'aide de `EvolSiteArbre::`*invFonctionChangementArbre*, on calcule la pseudo-inverse t_c de la fonction de survie en cette valeur u,
- si $t_c = -1$, alors :
  - si $\pi_i(t_c-) \neq \pi_i(t_c)$, alors le saut qui se produit à l'instant t_c est provoqué par un changement pour l'évolution π_i. On effectue donc un saut en t_c, vers le dinucléotide encodé z_c obtenu par `EvolSiteArbre::`*sautAGaucheAvecConditionnementMat*,
  - sinon, aucun changement ne se produit sur le reste de l'arête en cours. On définit z_0 comme l'état initial des deux arêtes issues de l'arête en cours (si elles existent).
- si $t_c > t_0$, alors on effectue un changement en l'instant t_c. La valeur du changement est régie par la méthode `EvolSiteArbre::`*quelChangementAvecConditionnementMat*.

Temps d'exécution : 47% pour `EvolSiteArbre::`*avancerUnChangementAvecConditionnement* dont :
- 29% pour les méthodes appelées `EvolSiteArbre::`*invFonctionChangementArbre*,
- 2% pour `EvolSiteArbre::`*sautAGaucheAvecConditionnementMat*,
- 6% pour `EvolSiteArbre::`*quelChangementAvecConditionnementMat*.

On décrit dans les paragraphes *Valeur après un saut provoqué par un changement de* π_i et *Valeur après un saut sans changement de* π_i les deux méthodes permettant de connaître la valeur du dinucléotide encodé à l'instant de changement. Chacune de ces méthodes nécessite la simulation d'une variable aléatoire.

Valeur après un saut provoqué par un changement de π_i.

La méthode `EvolSiteArbre::`*sautAGaucheAvecConditionnementMat* choisit la valeur du dinucléotide encodé après le saut effectué à l'instant t_c, conditionnellement à π_i, aux valeurs associées aux feuilles $z_i(T)$ de l'arbre et à la valeur z_0 du dinucléotide avant l'instant de changement.

Si $\pi_i(t_c-) = R$ et $\pi_i(t_c) = Y$, on choisit aléatoirement un élément z_c proportionnellement aux poids donnés par la fonction :

$$z_c \mapsto U_{R\to Y}(z_0, z_c)G_t(z_c, z_i(T)).$$

Si au contraire $\pi_i(t_c-) = Y$ et $\pi_i(t_c) = R$, une seule substitution est possible en l'instant t_c.

Valeur après un saut sans changement de π_i

La méthode `EvolSiteArbre::`*quelChangementAvecConditionnementMat* choisit la valeur du dinucléotide encodé après le saut effectué à l'instant t_c, conditionnellement à π_i, aux valeurs associées aux feuilles $z_i(T)$ de l'arbre et à la valeur z_0 du dinucléotide avant l'instant de changement.

Si $\pi_i(t_c) = R$, on choisit aléatoirement un élément z_c proportionnellement aux poids donnés par la fonction :

$$z_c \mapsto W_R(z_0, z_c)G_t(z_c, z_i(T)).$$

Si au contraire $\pi_i(t_c) = Y$, on choisit aléatoirement un élément z_c proportionnellement aux poids donnés par la fonction :

$$z_c \mapsto W_Y(z_0, z_c)G_t(z_c, z_i(T)).$$

9.3.4 Avancements particulaires pour un site

On suppose que l'évolution pour l'ensemble des particules $j \in [\![1, n]\!]$ est effectuée pour le site $i-1$. On cherche à mettre à jour les particules et à effectuer l'évolution pour le site i. On procède de la façon suivante :

1. On commence par calculer et enregistrer le poids de chaque particule. Pour toute particule j,
   (a) connaissant le dinucléotide encodé à la racine au site $i-1$, on calcule les probabilités d'obtenir le dinucléotide init à la racine au site i (vecteur de taille 3 ou 6) :
   $$P(r_i^j = \text{init} | r_{i-1}^j). \tag{9.2}$$
   La méthode correspondante est `EvolPartic::`*probaRacineSachantPasse*.
   (b) pour chaque dinucléotide encodé init, on calcule les probabilités :
   $$P(z_i(T) | z_{i-1}^j, r_i^j = \text{init}). \tag{9.3}$$
   La méthode correspondante est `EvolPartic::`*poidss*.
   (c) pour chaque dinucléotide encodé init, on effectue le produit des probabilités des équations (9.2) et (9.3) pour obtenir :
   $$P(r_i^j = \text{init}, z_i(T) | z_{i-1}^j). \tag{9.4}$$
   (d) On obtient l'ensemble des poids en sommant sur l'ensemble des valeurs initiales possibles init les probabilités de l'équation (9.4), c'est-à-dire :
   $$P(z_i(T) | z_{i-1}^j).$$
2. Si le rééchantillonnage des particules doit avoir lieu au site i, on rééchantillonne selon les différents poids obtenus. Pour cela, on utilise l'algorithme de Walker [117]. La méthode associée est `EvolPartic::`*reechParticules*.
3. En considérant les nouvelles particules, on définit la valeur du dinucléotide encodé r_i^j à la racine de l'évolution au site i. Pour cela, on remarque que pour une particule j fixée, $C := P(z_i(T) | z_{i-1}^j)$ est constante. L'équation (9.4) se réécrit donc :
   $$\text{init} \mapsto C \times P(r_i^j = \text{init} | z_{i-1}^j, z_i(T)),$$
   et il suffit de tirer aléatoirement une valeur init proportionnellement aux poids donnés par cette fonction. La méthode associée est `EvolPartic::`*avancerLettreInitiale*.
4. Enfin, on fait évoluer chaque particule le long de l'arbre pour le site considéré. La méthode associée est `EvolPartic::`*avancerSite* et consiste à répéter l'exécution de la méthode déjà décrite `EvolSiteArbre::`*avancerUnChangementAvecConditionnement* le nombre de fois nécessaire pour effectuer l'avancement complet d'un site le long de l'arbre.

Temps d'exécution : 97% pour `EvolPartic::`*avancerUnChangement* dont :
- 3% pour les méthodes appelées `EvolPartic::`*probaRacineSachantPasse*,
- 16% pour les méthodes appelées `EvolPartic::`*poidss*,
- 15% pour les méthodes appelées `EvolPartic::`*reechParticules*,
- 2% pour les méthodes appelées `EvolPartic::`*avancerLettreInitiale*,
- 59% pour les méthodes appelées `EvolPartic::`*avancerSite*.

Chapitre 10

Applications

Dans ce chapitre, on cherche à utiliser les estimateurs de la vraisemblance, par approximations markoviennes ou par approximations particulaires, présentés dans les chapitres 4 et 8.

Le chapitre est orienté autour de la résolution de trois problèmes usuels pour les chaînes de Markov cachées présentés dans l'introduction de Rabiner [99]. Ces trois problèmes sont les suivants.

1. Étant donné un modèle global d'évolution λ et des observations $\mathbf{y}$, on cherche à calculer la vraisemblance $P(\mathbf{y}|\lambda)$ des observations sous le modèle λ.
2. Étant donné un modèle global d'évolution λ et des observations $\mathbf{y}$, on cherche à trouver la *meilleure* séquence ancestrale permettant d'obtenir ces observations (où le sens de *meilleure* séquence est à définir).
3. Étant donné des observations $\mathbf{y}$, on cherche à obtenir le modèle d'évolution λ qui maximise la vraisemblance des observations $P(\mathbf{y}|\lambda)$.

Le cadre spécifique de notre modèle d'évolution (voir la chaîne de Markov cachée étudiée dans la section 7.3), en particulier le fait que l'espace d'états caché ne soit pas dénombrable, motive l'utilisation des estimateurs particulaires dans la résolution de ces problèmes.

Le travail réalisé dans ce chapitre pour résoudre ces problèmes fournit deux retombées majeures.

D'une part, on confirme numériquement les propriétés asymptotiques des estimateurs particulaires (théorème 8.2.3 et conjecture 8.2.6) pour calculer une approximation de la vraisemblance. Cela permet en outre de confronter les modèles d'évolution RN95+YpR avec d'autres modèles qui ne sont pas dans cette classe – par exemple le modèle GTR (défini dans la section 1.1), en effectuant une comparaison directe des vraisemblances (voir section 10.6).

D'autre part, les estimateurs particulaires ont permis de faire une recherche empirique systématique sur les paramètres du modèle d'évolution et les observations. On en déduit des gammes de paramètres pour lesquelles on met en évidence certains comportements non souhaitables lorsque l'on utilise les approximations markoviennes (mauvaise approximation de la vraisemblance, voir sections 10.1.2 et 10.2.1) ou lorsque l'on tronque les séquences d'observations (mauvaise inférence de la séquence ancestrale, voir section 10.4.3).

On sépare les modèles globaux d'évolution considérés dans ce chapitre en deux parties : les modèles typiques et les modèles atypiques (regroupés dans l'annexe A).

Les modèles typiques correspondent à des valeurs plausibles vis-à-vis des applications biologiques envisagées. Les paramètres associés à ces modèles sont soit issus d'estimations réalisées sur des séquences biologiques avec un logiciel de maximisation des paramètres (le logiciel `bppml` de Bio++ [15, 38, 39]), soit directement tirés uniformément sur un ensemble compact.

Les modèles atypiques signifient que l'on n'est plus dans une gamme de valeurs plausibles biologiquement, mais que l'on a choisie spécialement un modèle pour illustrer un comportement.

À ces modèles on associe également deux sortes de jeux de séquences observées : soit ces jeux sont simulés selon le modèle utilisé pour effectuer l'estimation (cas typique, simulés avec le logiciel Alfacinha [90]), soit on choisit spécifiquement des jeux de séquences.

On rappelle que trois approximations pour calculer ou approcher la vraisemblance ont été étudiées dans les chapitres précédents : le calcul direct (chapitre 3), l'approximation markovienne $\hat{L}_{k\text{-Markov}}$ (où $k = 1/2$ ou $k \geq 1$) et les approximations particulaires $\hat{L}_{n,r\text{-partic}}$ (où $n \geq 1$ et $r \geq 0$).

La première consiste à calculer directement les matrices de transitions pour en déduire la vraisemblance de la séquence observée (voir corollaire 3.5.1). On sait (section 3.5.1) que cette méthode n'est envisageable numériquement que pour les séquences de longueur faible (jusqu'à des séquences de longueur 6 ou 7 environ). Néanmoins, lorsqu'elle est accessible, elle fournit la valeur exacte de la vraisemblance.

Les approximations markoviennes négligent en chaque site les sites voisins au-delà d'un certain nombre de pas k fixé pour donner une approximation de la vraisemblance d'une séquence. Pour un nombre de pas k égale à 1 ou 2, cette méthode est numériquement utilisable sur des séquences longues (plus de 10000 sites). Par contre, on ne dispose pas à k fixé de contrôle de l'erreur commise par rapport à la valeur exacte de la vraisemblance. (voir section 4.2 pour une description de cette approximation)

Enfin, les approximations particulaires utilisent un ensemble de particules évoluant le long de la séquence avec ou sans rééchantillonnage. Ils donnent accès à une approximation de la vraisemblance des observations et de plus, des propriétés de consistance et de normalité asymptotique sont connues lorsque le nombre de particules utilisées tend vers l'infini (voir section 8.2.2). En particulier, ils donnent accès à une estimation de l'erreur commise sur les approximations effectuées.

On décrit maintenant le contenu des différentes sections du chapitre.

Dans les sections 10.1 et 10.2, on cherche à comparer la qualité d'estimation de ces trois estimateurs, pour des séquences courtes où le calcul de la vraisemblance exacte est accessible (section 10.1) ou non (section 10.2). On obtient que dans des cas atypiques, les approximations particulaires fournissent des estimations convergentes de la valeur exacte de la vraisemblance alors que l'approximation markovienne à un pas donne une estimation médiocre. Cela confirme la robustesse des estimateurs particulaires et le fait que l'erreur dans les estimateurs par approximation markovienne n'est pas contrôlée et que sa pertinence doit être vérifiée. Dans les cas typiques, l'estimation par approximation markovienne

à un pas donne en général une estimation de la vraisemblance aussi précise ou plus précise que celle par méthode particulaire utilisant un grand nombre de particules (100000 particules).

Dans la section 10.3, on étudie et valide les propriétés théoriques associées aux estimateurs particulaires de la log-vraisemblance. On vérifie tout d'abord la convergence puis la normalité asymptotique de ces estimateurs vers la log-vraisemblance. Enfin, on étudie le biais de ces estimateurs par rapport à la log-vraisemblance, en fonction du nombre de particules et de la longueur de séquence considérés.

On en conclut que ces estimateurs permettent d'approcher numériquement la log-vraisemblance de la séquence et répondent donc au premier problème de l'introduction de Rabiner [99] dans ce contexte.

Dans la section 10.4, on s'intéresse au deuxième problème de l'introduction de Rabiner [99] et on cherche à inférer la séquence ancestrale en calculant pour chaque site i le nucléotide ancestral le plus probable. On définit pour cela des méthodes utilisant les évolutions particulaires. Comme ces méthodes sont coûteuses en temps de calculs, on cherche à négliger les observations qui sont loin du site i considéré. On établit le nombre de sites voisins à prendre en compte autour de i pour inférer de façon pertinente le nucléotide ancestral en ce site. On propose enfin un algorithme permettant de reconstituer la séquence ancestrale complète.

Dans la section 10.5, on compare la viabilité des estimateurs de vraisemblances pour obtenir des estimations du maximum de vraisemblance. On compare en particulier les estimateurs consistants par triplets encodés (voir proposition 4.1.3) et particulaires (voir théorème 8.2.5). On illustre que les estimateurs particulaires sont coûteux en temps de calcul et qu'ils ne fournissent pas une estimation plus précise que celle par triplets encodés. Ainsi, on préfère se servir de l'estimateur par triplets encodés pour estimer le maximum de vraisemblance. On estime ensuite l'écart-type de l'estimateur du maximum de vraisemblance par triplets encodés en utilisant une méthode empirique et la méthode semi-empirique proposée dans la section 4.1.2.

Dans la section 10.6, on déduit des sections précédentes un algorithme permettant d'estimer (de façon consistante) la vraisemblance des observations pour une estimation (consistante) du modèle de maximum de vraisemblance. Cet algorithme utilise à la fois les estimateurs par triplets encodés et les estimateurs particulaires. Cela correspond à répondre au troisième problème de l'introduction de Rabiner [99].

On se sert de cet algorithme pour comparer des estimations de la vraisemblance au maximum de vraisemblance. On considère deux alignements de nucléotides et les classes de modèles d'évolution T92+CpGs, T92 et GTR. On obtient que pour les deux alignements, la prise en compte (à partir de la classe T92) du paramètre d'hypermutabilité CpG améliore environ dix fois plus la vraisemblance que la prise en compte des paramètres présents dans le modèle GTR.

10.1 Comparaison des approximations pour des séquences courtes

On regroupe dans cette section des exemples de modèles d'évolution associés à des jeux d'observations de longueurs courtes, c'est-à-dire pour lesquels le calcul de la vraisemblance exacte L est numériquement accessible. On cherche alors à comparer avec cette valeur exacte les différents estimateurs de vraisemblances utilisés : les estimateurs de vraisemblance $\hat{L}_{k\text{-Markov}}$ (avec $k \geq 1$) obtenus par approximation markovienne et les estimateurs de vraisemblance particulaires $\hat{L}_{n,r\text{-partic}}$ (où $n \geq 1$ et $r \in \{0, 1\}$) sans rééchantillonnage ou avec rééchantillonnage systématique.

On compare dans la section 10.1.1 la valeur exacte de la vraisemblance avec les estimations obtenues par approximations markoviennes. Pour cela, on choisit deux modèles d'évolution – λ_1 modèle atypique et λ_2 modèle typique – et on cherche à obtenir les pires séquences de longueur 4 et 5 vis-à-vis de l'approximation markovienne, c'est-à-dire les séquences pour lesquels la proportion d'écart de log-vraisemblance par rapport à la valeur exacte est la plus forte. On montre que les proportions d'écarts associées à ces séquences sont perceptibles dans les deux exemples et beaucoup plus fortes dans le cas du modèle atypique (de l'ordre de 10% dans le modèle atypique et de 10^{-4} dans le modèle typique). On en conclut que sur des séquences courtes, l'estimation par approximation markovienne donne une approximation plausible de la log-vraisemblance pour certains modèles typiques, mais que cette approximation est moins pertinente pour certains modèles atypiques.

On compare ensuite dans la section 10.1.2 les estimations obtenues par approximations markoviennes et par les méthodes particulaires par rapport à la valeur exacte de log-vraisemblance. On prend de nouveau les modèles d'évolution λ_1 et λ_2 et on leur associe des observations de longueur 4, 5 et 6, pour lesquels le calcul de la vraisemblance exacte est possible.

On obtient comme attendu que les méthodes particulaires fournissent des estimations convergentes (quand le nombre de particules croît) de la valeur exacte de log-vraisemblance, même dans les cas où l'approximation markovienne à un pas donne une estimation médiocre.

Pour le modèle typique, on observe que l'estimation par approximation markovienne à un pas donne une estimation de la vraisemblance plus précise qu'avec les estimations par méthodes particulaires utilisant 100000 particules, pour un temps de calcul informatique beaucoup plus faible. Néanmoins, pour le modèle atypique, il suffit de 100 particules pour obtenir une estimation aussi précise que celle utilisant l'approximation markovienne à un pas. Ainsi, les deux méthodes d'estimations apparaissent complémentaires dans la recherche efficace de la log-vraisemblance.

10.1.1 Approximation markovienne à un pas et valeur exacte

On étudie la qualité de l'estimation par approximation markovienne à un pas sur deux modèles en comparant sur des jeux d'observations de longueur 4 et 5 la proportion d'écart entre la valeur exacte et l'estimation effectuée. On cherche alors pour ces deux modèles les séquences de longueurs 4 et 5 pour lesquels la proportion d'écart de log-vraisemblance est la plus forte.

Données. On cherche à définir un modèle atypique et un modèle typique. Pour le modèle atypique, on s'inspire du modèle de l'exemple limite 5.1.10 pour définir le modèle global de l'exemple 10.1.1. Pour le modèle typique, on choisit de tirer uniformément tous les paramètres d'évolution du modèle RN95+YpR dans l'intervalle $[0, 10]$ et on obtient le modèle de l'exemple 10.1.2.

Exemple 10.1.1. *(cas atypique). On considère le modèle d'évolution* M_3 *défini par (où les coefficients non indiqués sont égaux à* 0.01*) :*

$$v_G = 10,\ r_{CG\to CA} = 100.$$

On considère l'arbre $\mathsf{T}_2(1)$ *constitué de deux arêtes de même longueur* 1 *(voir annexe A). On choisit la loi à la racine* $\mathsf{R}_{\mathsf{M}_3}$ *associée à la loi stationnaire du modèle* M_3*.*

On note le modèle complet $\lambda_1 = (\mathsf{R}_{\mathsf{M}_3}, \mathsf{T}_2(1), \mathsf{M}_3)$*.*

Exemple 10.1.2. *(cas typique). On considère le modèle d'évolution* M_4 *donné par :*

$$\begin{aligned}
&v_A = 7.199,\ v_C = 6.235,\ v_G = 0.241,\ v_T = 7.313,\\
&w_A = 8.702,\ w_C = 6.914,\ w_G = 7.538,\ w_T = 0.314,\\
&r_{CG\to CA} = 3.821,\ r_{CA\to CG} = 3.363,\ r_{TA\to TG} = 3.340,\ r_{TG\to TA} = 2.517,\\
&r_{CA\to TA} = 5.614,\ r_{CG\to TG} = 8.155,\ r_{TA\to CA} = 8.020,\ r_{TG\to CG} = 7.705.
\end{aligned}$$

On lui associe l'arbre $\mathsf{T}_2(1)$ *et la loi à la racine* $\mathsf{R}_{\mathsf{M}_4}$ *associée à la loi stationnaire du modèle* M_4*.*

On note le modèle complet $\lambda_2 = (\mathsf{R}_{\mathsf{M}_4}, \mathsf{T}_2(1), \mathsf{M}_4)$*.*

On choisit comme observations associées à ces modèles les séquences identiques sur chacune des deux feuilles pour lesquelles la vraisemblance ne peut pas être découpée en deux morceaux distincts au sens du corollaire 3.5.2. Explicitement, pour une séquence Φ-encodée s, on dit que s est minimale si la séquence π-encodée $\pi(s)$ associée ne contient pas la suite de lettres RY. Puis pour $m \in \{4, 5\}$, on note $S_{m,0}$ l'ensemble des séquences Φ-encodées de longueur m minimales et on considère l'ensemble des observations $\mathbf{Y}_m$ données par $\mathbf{y} = (s, s)$ avec $s \in S_{m,0}$.

Méthode. Pour $m \in \{4, 5\}$, pour chacun des deux exemples 10.1.1 et 10.1.2, pour l'ensemble des observations de $\mathbf{Y}_m$, on calcule l'estimation de la log-vraisemblance par approximation markovienne à un pas $\hat{L}_{1\text{-Markov}}$ ainsi que la valeur de log-vraisemblance exacte L.

Pour $m \in \{4, 5\}$ et chacun des deux exemples, on représente sur la figure 10.1 le diagramme en boîte associé aux proportions d'écart $\frac{L-\hat{L}_{1\text{-Markov}}}{L}$ de log-vraisemblance entre la valeur estimée par approximation markovienne à un pas et la valeur exacte. On déduit également dans chacun des cas les séquences dont l'écart entre l'estimation par approximation markovienne et la valeur exacte est le plus fort.

Résultats. On observe dans les quatre cas que pour la plupart des séquences, l'écart entre l'estimation par approximation markovienne à un pas et la valeur exacte est faible, dans le sens où l'écart interquartile de la proportion d'écart est plus petite que trois pour mille.

Pour les séquences menant à des proportions d'écart extrêmes, on observe une différence de comportement entre les exemples typiques et atypiques. Pour les deux cas typiques, la

proportion d'écart reste faible et l'estimation par approximation markovienne donne une approximation plausible de la log-vraisemblance, avec dans le pire des cas une proportion d'écart de $2.78.10^{-4}$ pour la séquence $TTTY$ et de $-2.63.10^{-4}$ pour la séquence $CTTCY$.

Pour les deux cas atypiques, la proportion d'écart peut être importante, avec dans le pire des cas une proportion d'écart de 8.60% pour la séquence $RAAA$ et de 15.0% pour la séquence $RAAAA$.

10.1.2 Approximations particulaires et markoviennes

Sur des séquences courtes, on utilise à la fois les estimations par approximations markoviennes et celles utilisant les méthodes particulaires dans le but de comparer la qualité des valeurs obtenues par rapport à la valeur exacte.

Données. On reprend les modèles complets λ_1 et λ_2 des exemples 10.1.1 et 10.1.2. On associe au modèle λ_1 atypique les séquences observées

$$(RAAA, RAAA), (RAAAA, RAAAA) \text{ et } (RAAAAA, RAAAAA)$$

et au modèle λ_2 typique les séquences

$$(TTTA, TTTA), (TTTAA, TTTAA) \text{ et } (TTTAAA, TTTAAA),$$

respectivement de longueur 4, 5 et 6.

Méthode. Pour les modèles λ_1 et λ_2 et les séquences associées de longueur $m \in \{4, 5, 6\}$, on calcule les estimations de log-vraisemblances données par :

- Pour $n \in \{100, 1000, 10000, 100000\}$ et avec 100 répétitions :
  - $\hat{L}_{n,\text{0-partic}}$ méthode particulaire sans rééchantillonnage.
  - $\hat{L}_{n,\text{1-partic}}$ méthode particulaire avec rééchantillonnage à chaque pas.
- $\hat{L}_{k\text{-Markov}}$ avec $k \in [\![1, m-3]\!]$.

D'autre part, on calcule la log-vraisemblance exacte.

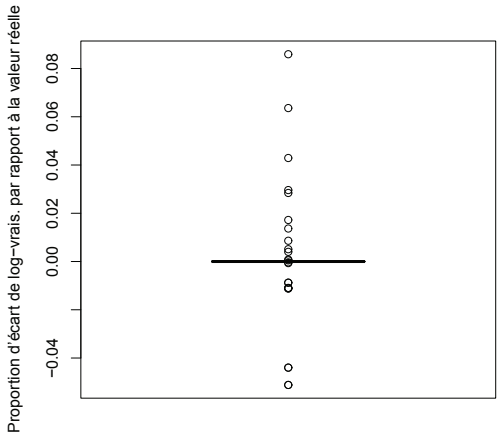
On représente ensuite sur la figure 10.2 les estimations de log-vraisemblance issues des différents estimateurs ainsi que la valeur exacte de log-vraisemblance. Sur la figure 10.3, on trace les écarts quadratiques moyens empiriques associés aux estimations obtenues.

Résultats.

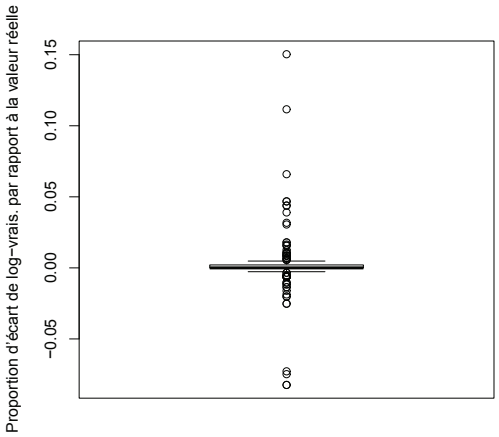
Convergence des estimateurs particulaires quand le nombre de particules tend vers l'infini. Avec et sans rééchantillonnage et pour tous les exemples, on observe comme attendu une convergence à vitesse $\frac{1}{\sqrt{n}}$ vers la valeur exacte (voir les représentations des écarts quadratiques moyens empiriques sur la figure 10.3).

Une étude dans un contexte plus général de la convergence numérique des estimateurs particulaires est donnée dans la section 10.3.2. En particulier, le présence et l'estimation du biais est considérée dans la section 10.3.3 (que l'on observe par exemple sur les diagrammes en boîtes de la figure 10.2, pour le modèle atypique pour la longueur 6 sans rééchantillonnage).

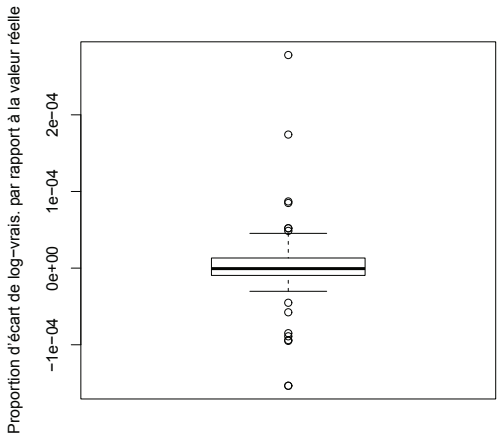
Également, les différences de comportements et de précision avec et sans rééchantillonnage sont étudiées spécifiquement dans la section 10.2.3.



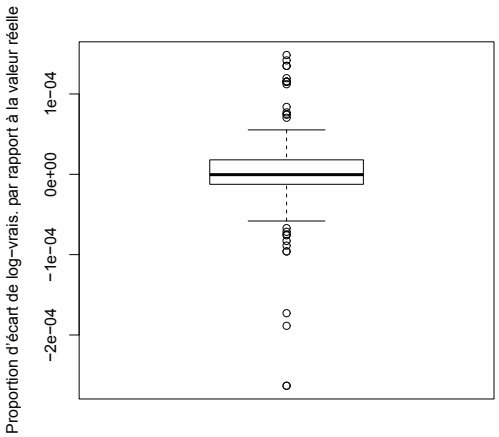
Modèle atypique. Longueur 4.



Modèle atypique. Longueur 5.



Modèle typique. Longueur 4.



Modèle typique. Longueur 5.

FIGURE 10.1 – Diagrammes en boîte associés aux proportions d'écart de log-vraisemblance $\frac{L-\hat{L}_{1\text{-Markov}}}{L}$ entre la valeur estimée par approximation markovienne à un pas et la valeur exacte, pour l'exemple atypique 10.1.1 et l'exemple typique 10.1.2.

Comparaisons des estimations. On compare les différentes estimations à l'aide de la figure 10.3. On observe de façon générale que les estimations sont plus précises dans le modèle typique que dans le modèle atypique. De plus, on observe qu'il suffit de 100 particules pour obtenir une estimation aussi précise que celle utilisant l'approximation markovienne à un pas dans le cas atypique alors qu'il en faut 100000 dans le cas typique présenté.

10.2 Comparaison des estimateurs de vraisemblance

Dans la section 10.1.2, on a comparé sur deux modèles les estimateurs par approximations markoviennes et par les méthodes particulaires, en utilisant des séquences courtes pour pouvoir effectuer le calcul exact de la vraisemblance. En complément, on cherche ici à comparer ces estimateurs sur des séquences de longueur plus élevée, pour lesquelles le calcul de la vraisemblance exacte n'est plus possible numériquement.

Tout d'abord, on utilise dans la section 10.2.1 trois modèles atypiques sur lesquels on met en évidence la convergence de l'estimateur particulaire avec rééchantillonnage et l'insuffisance de l'estimateur par approximation markovienne à un pas sur ces modèles, comme on pouvait s'y attendre d'après l'exemple atypique présenté dans la section 10.1.2. On observe de plus que les estimations obtenues par approximations markoviennes à trois pas sont plus précises que celles issues de l'exemple limite 5.1.10.

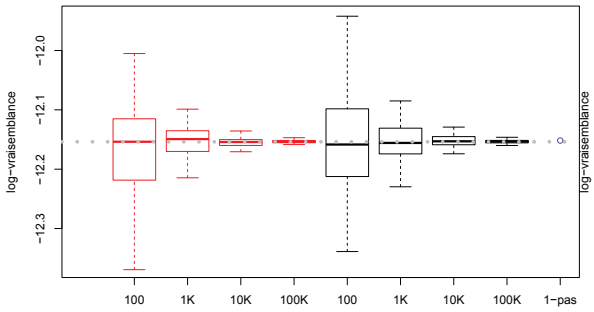
Ensuite, on compare dans la section 10.2.2 sur un ensemble de modèles d'évolution typiques la proportion d'écart de log-vraisemblance entre une valeur de référence et les estimations particulaires ou les estimations par approximations markoviennes. On observe que pour ces exemples typiques, les estimations obtenues par approximations markoviennes sont en général proches de celles obtenues par les méthodes particulaires.

Enfin, dans la section 10.2.3, on regroupe les commentaires concernant les différences de comportement entre les estimations utilisant les méthodes particulaires avec rééchantillonnage à chaque pas ou sans rééchantillonnage. On compare la convergence, la précision et le coût numérique de ces méthodes à nombre de particules identiques. On en conclut que les méthodes particulaires avec rééchantillonnage systématique sont préférées dans le but d'éviter les problèmes de dégénérescence des poids.

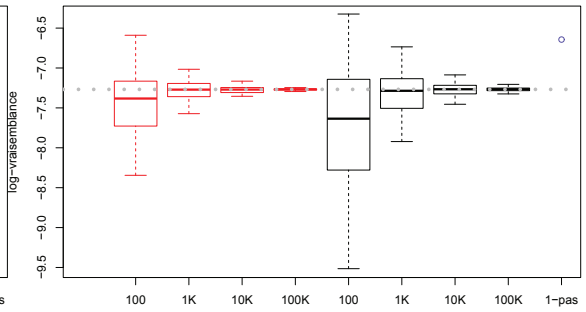
10.2.1 Approximations particulaires et markoviennes : cas atypiques

Sur des séquences de longueur 10 et 100, on va observer sur des modèles atypiques les valeurs estimées de la log-vraisemblance obtenues par les estimateurs par approximations markoviennes ainsi que par les estimateurs particulaires.

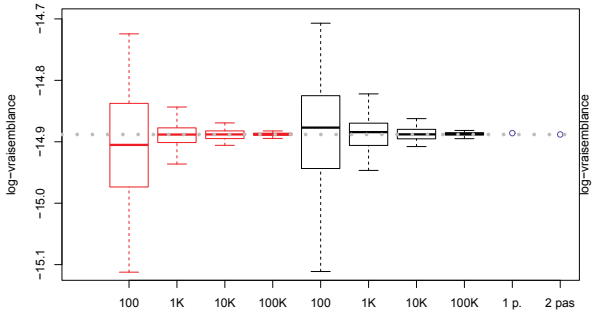
Données. On considère dans l'exemple 10.2.1 deux modèles atypiques proches du modèle λ_1 (défini dans l'exemple 10.1.1), qui n'en diffère que de part la loi à la racine choisie et la longueur des branches de l'arbre considéré. Le modèle atypique de l'exemple 10.2.2 est quant à lui choisi à partir du modèle d'évolution $\mathsf{M}_{extIrr}(\varepsilon)$ défini dans la section 5.2, en choisissant $\varepsilon = 0.01$.



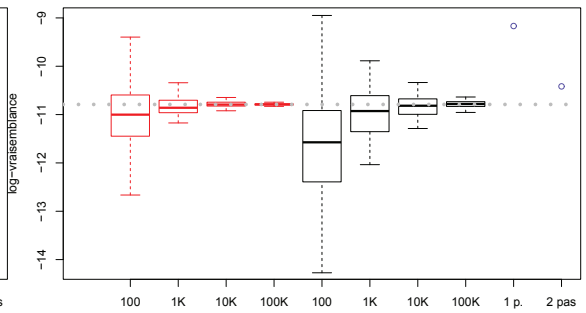
Modèle typique. Longueur 4.



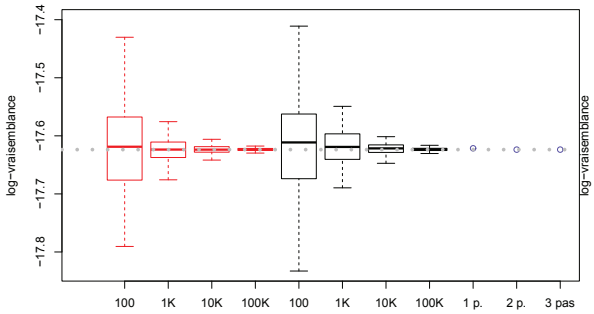
Modèle atypique. Longueur 4.



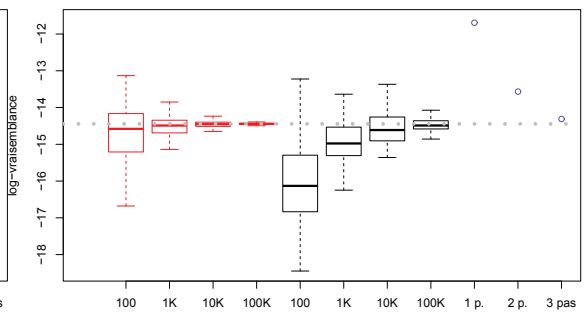
Modèle typique. Longueur 5.



Modèle atypique. Longueur 5.

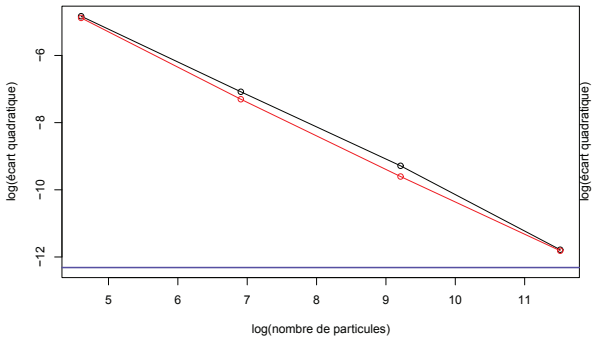


Modèle typique. Longueur 6.

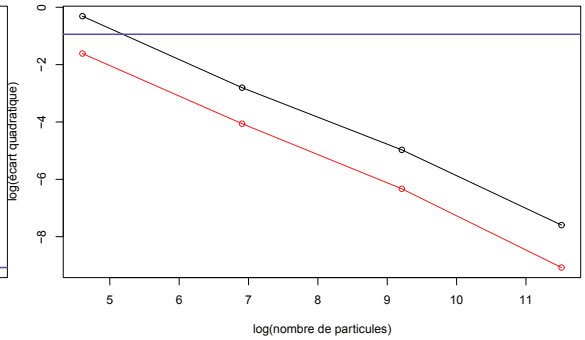


Modèle atypique. Longueur 6.

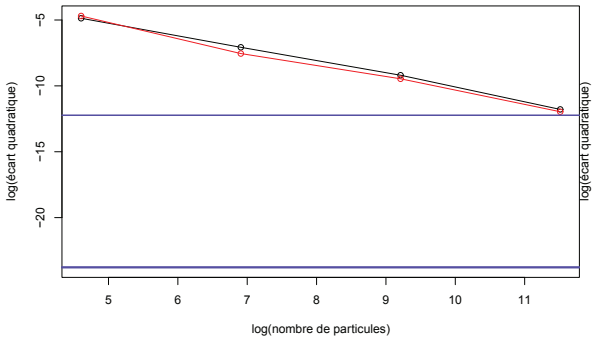
FIGURE 10.2 – Pour les modèles et les observations associées de la section 10.1.2, représentation de l'estimation particulaire sans rééchantillonnage (en noir), avec rééchantillonnage systématique (en rouge), pour $n \in \{100, 1000, 10000, 100000\}$ particules, et représentation en bleu de l'estimation par approximation markovienne à $k \in \{1, 2, 3\}$ pas. La ligne pointillée grise correspond à la valeur de log-vraisemblance exacte.



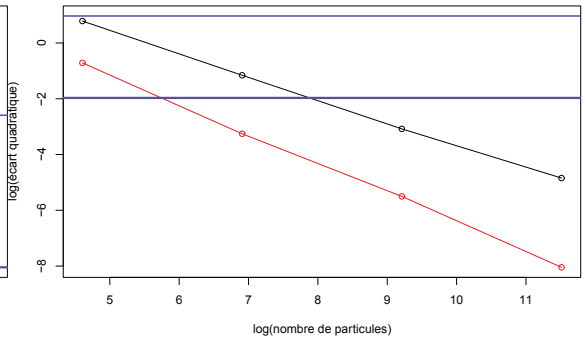
Modèle typique. Longueur 4.



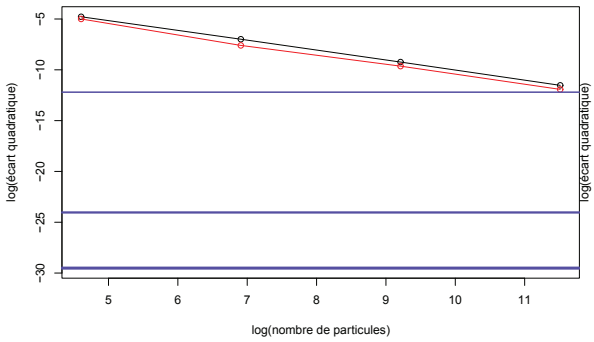
Modèle atypique. Longueur 4.



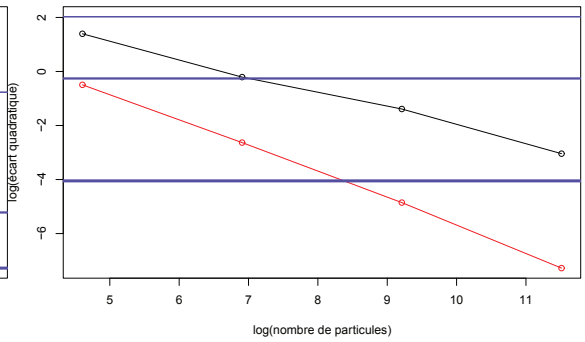
Modèle typique. Longueur 5.



Modèle atypique. Longueur 5.



Modèle typique. Longueur 6.



Modèle atypique. Longueur 6.

FIGURE 10.3 – Pour les modèles et les observations associées de la section 10.1.2, évolution de l'écart quadratique moyen empirique de l'estimation particulaire sans rééchantillonnage (en noir) et avec rééchantillonnage systématique (en rouge) en fonction du nombre de particules, sur un repère log-log. Les lignes bleues représentent les écarts quadratiques entre la valeur exacte et les estimations par approximations markoviennes (plus la ligne est épaisse et plus le nombre de pas est important).

Exemple 10.2.1. *(cas atypique). On rappelle que le modèle d'évolution* M_3 *est défini par (les coefficients non indiqués sont égaux à* 0.01*) :*

$$v_G = 10,\ r_{CG\to CA} = 100.$$

Pour $t \in \{0.1, 1\}$*, on considère l'arbre* $\mathsf{T}_2(t)$ *constitué de deux arêtes de même longueur* t*. On choisit une loi à la racine* R^C *fixe, égale au nucléotide* C *en chaque site.*

On note le modèle complet $\lambda_3(t) = (R^C, \mathsf{T}_2(t), \mathsf{M}_3)$*.*

Exemple 10.2.2. *(cas atypique). On fixe* $\varepsilon = 0.01$*. On rappelle que le modèle d'évolution* $\mathsf{M}_{extIrr}(\varepsilon)$ *est donné par (les coefficients non indiqués sont nuls) :*

$$v_C = v_G = 1, \quad r_{TG\to CG} = r_{CG\to CA} = r_{CA\to TA} = 1/\varepsilon.$$

$$v_A = v_T = w_A = w_C = w_G = w_T = \varepsilon.$$

On considère l'arbre $\mathsf{T}_2(10)$ *et la loi à la racine donnée par l'approximation markovienne à un pas de la loi stationnaire du modèle* $\mathsf{M}_{extIrr}(\varepsilon)$ *(voir annexe A).*

On note λ_4 *le modèle complet obtenu.*

Pour un choix du nombre de sites $m \in \{10, 100\}$, on associe pour tous les modèles les observations données par deux séquences observées identiques :

$$G\underbrace{A\ldots A}_{m-1}.$$

Méthode. On utilise pour les modèles $\lambda_3(0.1)$, $\lambda_3(1)$ et λ_4 et pour $m \in \{10, 100\}$ les estimations de la log-vraisemblances données par respectivement :

- Pour $n \in \{100, 1000, 10000, 100000\}$ et avec 100 répétitions :
  - $\hat{L}_{n,\text{0-partic}}$ méthode particulaire sans rééchantillonnage.
  - $\hat{L}_{n,\text{1-partic}}$ méthode particulaire avec rééchantillonnage à chaque pas.
- $\hat{L}_{k\text{-Markov}}$ avec $k = 1, 2, 3$.

On représente sur les six graphiques de la figure 10.4 les estimations de log-vraisemblance obtenus et on choisit comme valeur de référence la moyenne des 100 estimations issus de $\hat{L}_{100000,\text{1-partic}}$.

Résultats.

Convergence des estimateurs particulaires. Sur la figure 10.4, la convergence semble apparaître pour les estimations particulaires avec rééchantillonnage dans tous les exemples. Par contre, l'estimation particulaire sans rééchantillonnage ne semble cohérente qu'avec un choix de $m = 10$ sites. Pour le choix de 100 sites, les estimations ne sont pas correctes. Les différences de comportements et de précision avec et sans rééchantillonnage sont commentées dans la section 10.2.3.

Qualité des estimations par approximations markoviennes. Pour les modèles atypiques $\lambda_3(0.1)$ et $\lambda_3(1)$, on observe une différence importante entre l'estimation par approximation markovienne à 1, 2 et 3 pas. Par contre, on observe une différence beaucoup plus faible que pour le modèle limite entre l'estimation par approximation markovienne à 3 pas et l'estimation particulaire. Cela s'explique par le choix des paramètres de taux de sauts, tous supérieurs ou égaux à 0.01 : pour obtenir les observations voulues, les évolutions incluant des substitutions v_A ou v_C vont contribuer à la vraisemblance, alors que cela n'est pas le cas dans le modèle limite qui s'appuie exclusivement sur une chaîne de dépendance reliée à une permutation fixée.

Pour le modèle atypique λ_4, on observe de nouveau une différence importante entre l'estimation par approximation markovienne à 1, 2 et 3 pas.

10.2.2 Approximations particulaires et markoviennes : cas typiques

Sur des séquences de longueur 10 et 100, on compare sur des modèles typiques les valeurs estimées de la log-vraisemblance obtenues par les estimateurs par approximations markoviennes et par estimateurs particulaires.

Données On considère 100 modèles d'évolution RN95+YpR $(\mathsf{M}^l)_{l\in[\![1,100]\!]}$ dont chacun des 16 paramètres est tiré uniformément dans $[0, 10]$. Pour chaque modèle d'évolution M^l, on choisit l'arbre $\mathsf{T}_2(1)$ et la loi à la racine donnée par l'approximation markovienne à un pas de la loi stationnaire du modèle M^l. On obtient alors 100 modèles complets $(\lambda^l)_{l\in[\![1,100]\!]}$.

On associe à chacun des ces modèles quatre jeux séquences, donnés par les couples de séquences suivants :

$$AAAAAAAAAA \qquad AAAAAAAAAA, \tag{10.1}$$

$$CTTTCCGGGG \qquad CTCCTGAAGG, \tag{10.2}$$

$$\underbrace{A\dots A}_{100} \qquad \underbrace{A\dots A}_{100} \tag{10.3}$$

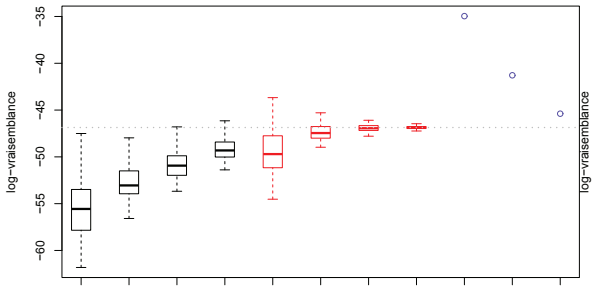
Le dernier couple est constitué de deux séquences de longueurs 100, chacune de la forme $Y\dots YR\dots R$ une fois π-encodée :

$$TTCTC\dots AGAAG \qquad TTTTC\dots AGAAA \tag{10.4}$$

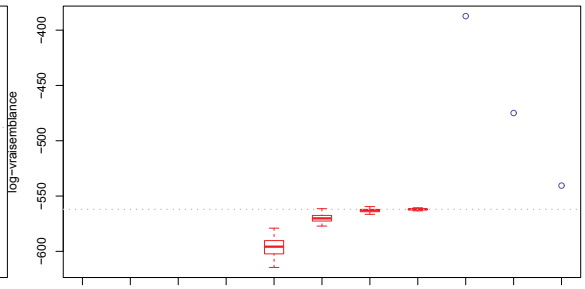
On remarque que l'on a choisi ces observations de manière à ce qu'il n'existe pas de couple de sites $(i, i+1)$ qui une fois Φ-encodé est égal à :

$$RY \qquad RY.$$

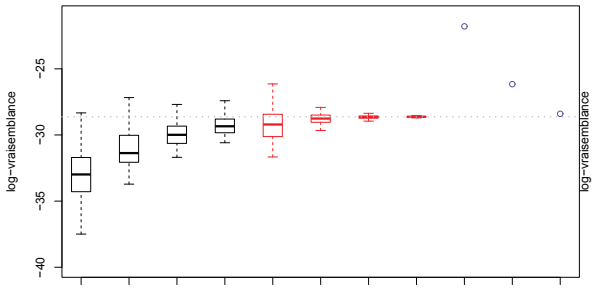
Si un tel couple de sites existait, le corollaire 3.5.2 de découpage RY permettrait de se ramener à deux jeux de séquences de longueurs inférieures.



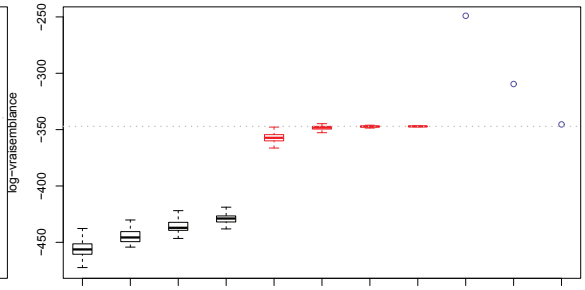
Cas atypique, modèle $\lambda_3(0.1)$ et $m = 10$.



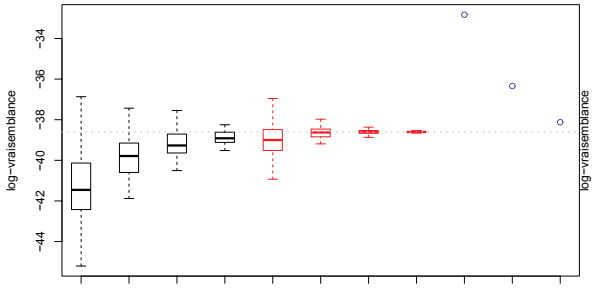
Cas atypique, modèle $\lambda_3(0.1)$ et $m = 100$.



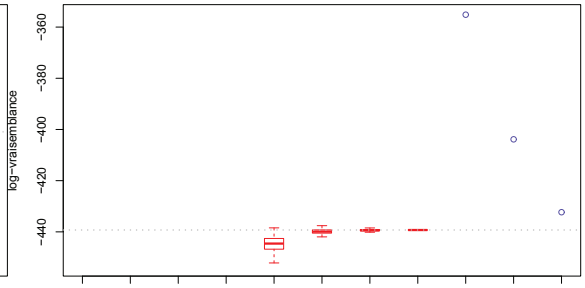
Cas atypique, modèle $\lambda_3(1)$ et $m = 10$.



Cas atypique, modèle $\lambda_3(1)$ et $m = 100$.



Cas atypique, modèle λ_4 et $m = 10$.



Cas atypique, modèle λ_4 et $m = 100$.

FIGURE 10.4 – Pour les modèles atypiques de la section 10.2.1, représentation de l'estimation particulaire sans rééchantillonnage (en noir), avec rééchantillonnage systématique (en rouge), pour $n \in \{100, 1000, 10000, 100000\}$, et représentation en bleu de l'estimation par approximation markovienne, pour $k \in \{1, 2, 3\}$ pas. La ligne pointillée grise correspond à la moyenne des 100 estimations issues de $\hat{L}_{100000,\text{1-partic}}$.

Méthode. Pour chacun des quatre jeux d'observations, pour chaque modèle, on calcule les estimations de la log-vraisemblance données par respectivement :

- Pour $n \in \{100, 1000\}$ et avec 100 répétitions :
  - $\hat{L}_{n,\text{0-partic}}$ méthode particulaire sans rééchantillonnage.
  - $\hat{L}_{n,\text{1-partic}}$ méthode particulaire avec rééchantillonnage à chaque pas.
- $\hat{L}_{k\text{-Markov}}$ avec $k = 1, 2, 3$.

On calcule ensuite pour chaque jeux d'observations et chaque modèle les écarts entre les estimations obtenues et la valeur de référence C donnée par la moyenne des 100 estimations issues de $\hat{L}_{1000,\text{1-partic}}$. On représente sur les graphiques de la figure 10.5 les différentes proportions d'écarts $\frac{C-\hat{L}}{C}$ pour chacune des estimations $\hat{L}$.

Résultats. On remarque que les écarts entre les valeurs de référence et les estimations par approximations markoviennes sont faibles, inférieurs à 1%, à l'exception de l'écart obtenu pour le modèle 77 pour les feuilles (10.1) et (10.3) où il atteint environ 3%. Hormis pour ce modèle, les écarts obtenus avec les estimations par approximation markovienne à deux ou trois pas sont plus faibles ou du même ordre de grandeur que les écarts entre les valeurs de référence et les estimations par méthodes particulaires $\hat{L}_{100,\text{0-partic}}$, $\hat{L}_{1000,\text{0-partic}}$ et $\hat{L}_{100,\text{1-partic}}$.

Ainsi, dans la plupart des cas les estimations par approximation markovienne fournissent des estimations plausibles et proches de celles obtenues par les estimateurs particulaires.

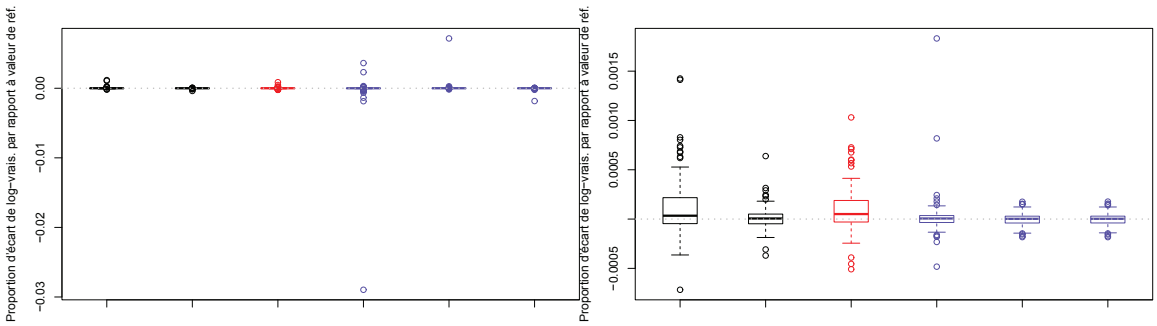
Autres cas typiques. On a présenté dans cette section une comparaison entre l'estimation par approximations markoviennes et par méthodes particulaires pour des cas typiques tirés aléatoirement. On compare dans la section 10.6 les résultats obtenus sur deux alignements de séquences génomiques. On observe (voir les figures 10.30 et 10.31) une différence non quantifiable entre les deux méthodes d'estimations, dès l'approximation markovienne à un pas.

10.2.3 Approximations particulaires avec et sans rééchantillonnage

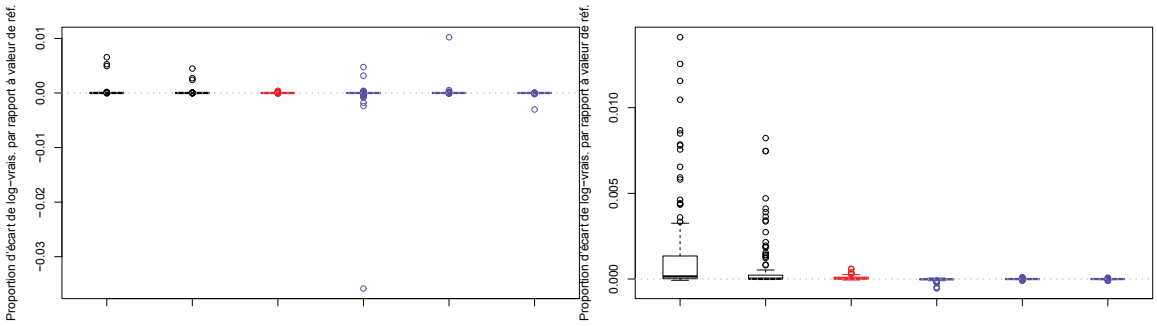
On reprend les différentes simulations effectuées dans les sections 10.1 et 10.2, dans le but de comparer spécifiquement les différences de comportement entre les estimations particulaires avec rééchantillonnage systématique et sans rééchantillonnage.

Convergence. Dans tous les exemples testés, constitués de modèles d'évolution typiques et atypiques et de séquences de longueur plus petite que 100 nucléotides, les estimateurs particulaires avec rééchantillonnage systématique sont convergents numériquement (voir figures 10.2 et 10.4).

Par contre, la convergence numérique des estimateurs particulaires sans rééchantillonnage n'est vérifiée que sur des séquences courtes (de longueur 4 à 6, voir figure 10.2) ou pour des modèles d'évolution typiques (voir figure 10.5). Sur des modèles atypiques associés à des séquences de longueur 100, les estimations obtenues sont erronées ou n'aboutissent pas (pour les graphiques de la figure 10.4 où les diagrammes en boîte sans rééchantillonnage ne sont pas indiqués, les sorties du programme pour les estimations de log-vraisemblance sont $-\infty$).



Modèle typique, $m = 10$, feuilles (10.1). Modèle typique, $m = 10$, feuilles (10.2).



Modèle typique, $m = 100$, feuilles (10.3). Modèle typique, $m = 100$, feuilles (10.4).

FIGURE 10.5 – Pour les modèles typiques de la section 10.2.2, représentation des proportions écarts entre la moyenne des 100 estimations issues de l'estimateur $\hat{L}_{1000,\text{1-partic}}$ et de l'estimation particulaire sans rééchantillonnage pour $n \in \{100, 1000\}$ (en noir), de l'estimation particulaire avec rééchantillonnage systématique pour $n = 100$ (en rouge) et de l'estimation par approximation markovienne, pour $k \in \{1, 2, 3\}$ (en bleu). La ligne pointillée grise correspond à la constante nulle.

Les différences de qualité observées entre la méthode sans rééchantillonnage et avec rééchantillonnage ne sont pas surprenantes d'après le comportement dégénérescent de la méthode sans rééchantillonnage quand m augmente (voir section 8.1.5).

Précision et coût. On observe dans tous les exemples (voir figures 10.3 et 10.5) que la méthode avec rééchantillonnage fournit une estimation plus précise que l'estimation sans rééchantillonnage à nombre de particules égal, avec en particulier une variance plus faible et un biais moins présent. Néanmoins, la méthode avec rééchantillonnage nécessite une étape supplémentaire de calculs et est plus coûteuse en temps à nombre de particules fixés.

Sur trois des exemples étudiés, on compare le temps de calcul nécessaire pour obtenir 100 répétitions d'estimations de log-vraisemblance à l'aide des méthodes particulaires $\hat{L}_{n,r\text{-partic}}$ avec $(r = 1)$ et sans $(r = 0)$ rééchantillonnage, pour $n \in \{100, 1000\}$. Les exemples considérés sont les suivants.

- Test (A) : figure 10.2, modèle typique λ_2, séquences de longueur 5.
- Test (B) : figure 10.4, modèle atypique $\lambda_3(1)$, séquences de longueur 10.
- Test (C) : figure 10.4, modèle atypique $\lambda_3(1)$, séquences de longueur 100.

On effectue les calculs sur un processeur cadencé à 3.3 GHz et on obtient les temps suivants, exprimés en secondes.

	(A)	(B)	(C)
$r = 0, n = 100$	93	9	92
$r = 1, n = 100$	99	10	120
$r = 0, n = 1000$	919	87	1018
$r = 1, n = 1000$	967	102	1340

En termes d'augmentation pour les estimations avec rééchantillonnages par rapport à ceux sans rééchantillonnage, on obtient :

	(A)	(B)	(C)
$n = 100$	6%	11%	30%
$n = 1000$	5%	17%	32%

On observe sur ces trois exemples que le coût de calcul supplémentaire croît en fonction du nombre de sites considérés et peu en fonction du nombre de particules utilisées. Toutefois, le coût à consentir reste acceptable même pour le test (C) utilisant des séquences de longueur 100.

Globalement, on conclut que les méthodes particulaires avec rééchantillonnage systématique sont préférées par rapport à celles sans rééchantillonnage, car plus robustes vis-à-vis des modèles atypiques ou associés à des séquences longues.

10.3 Fluctuations des approximations issues des méthodes particulaires

Cette section est consacrée à l'étude des fluctuations des approximations issues des méthodes particulaires avec rééchantillonnage systématique. On vérifie d'abord les propriétés théoriques de convergence et de normalité asymptotique des estimateurs associés aux rapports de vraisemblance $p(z_{i+1}(T) \mid z_{1:i}(T))$ ainsi que des estimateurs $\hat{L}_{n,\text{1-partic}}$ de la log-vraisemblance, avant d'exhiber et de quantifier le biais entre la log-vraisemblance et les estimations particulaires associées.

Dans la section 10.3.1, on cherche à vérifier le théorème 8.1.8. D'après ce théorème, on sait pour chaque site $i \in [\![1, m]\!]$, l'estimateur particulaire $\hat{\theta}^i(n)$ à n particules de $\theta^i = p(z_{i+1}(T)|z_{1:i}(T))$ vérifie que

$$\sqrt{n}\left(\hat{\theta}_i^n - \theta_i\right)$$

converge en loi vers une loi normale centrée d'écart-type fixé $\sigma^i_{\text{APF optimal}}$.

On teste alors cette propriété théorique sur trois modèles globaux d'évolution – deux modèles typiques et un modèle atypique – et des observations associées de longueur 100. On vérifie tout d'abord la convergence de ces estimateurs en introduisant un pseudo écart quadratique permettant d'estimer l'écart quadratique moyen. On en déduit comme attendu une décroissance en $1/n$ de ce pseudo écart quadratique en fonction du nombre de particules n.

Ensuite, on vérifie la normalité des échantillons par le test de Shapiro-Wilk [104]. On obtient que l'hypothèse de normalité ne peut pas être rejetée avec un risque à 5%, même dans l'exemple atypique. On estime aussi les écarts-types théoriques $\sigma^i_{\text{APF optimal}}$ pour chaque site i.

Dans le dernier paragraphe de la section 10.3.1, on mesure les corrélations à un ou plusieurs pas des estimations associées aux trois exemples considérés. En effet, on sait que pour deux sites $i \neq i'$, les estimateurs $\hat{\theta}^i(n)$ et $\hat{\theta}^{i'}(n)$ ne sont pas indépendants en général et on cherche à quantifier le nombre de pas pour lequel les corrélations deviennent non significatives. On observe une corrélation à un pas forte dans le cas de l'exemple atypique, qui diminue lorsque l'on augmente le nombre de pas jusqu'à devenir non significative à partir de 8 pas. Pour les deux modèles typiques, on observe que la corrélation n'est pas significative, même à un pas.

Dans la section 10.3.2, on étudie la validité du théorème 8.2.5 lorsque le nombre de particules n est grand devant le nombre de sites considérés m. Ce théorème indique que l'estimateur $\hat{L}_{n,r\text{-partic}}$ de la log-vraisemblance $\log p(z_{1:m}(T))$ est consistant et asymptotiquement normal.

Avec une méthode similaire à celle décrite dans la section 10.3.1, on vérifie sur deux exemples la convergence des estimateurs en montrant une décroissance en $1/n$ d'un pseudo écart quadratique en fonction du nombre de particules n à partir d'un certain nombre de particules. De même, on montre que l'hypothèse de normalité ne peut pas être rejetée avec un risque à 5%.

Lorsque le nombre de particules n est du même ordre de grandeur ou plus petit que le nombre de sites m, le théorème asymptotique 8.2.5 ne s'applique pas. On teste dans ce cas la présence d'un biais dans l'estimation de la log-vraisemblance. D'après la conjecture 8.2.6, la forme du biais proposée quand le nombre de particules n est du même ordre de grandeur que le nombre de sites m est proportionnel à m/n. On considère alors dans la section 10.3.3 deux modèles descriptifs pour exprimer la log-vraisemblance en fonction des estimateurs particulaires obtenues. Le premier modèle est sans biais alors que le second contient un terme de biais proportionnel à m/n. Sur un exemple constitué d'un modèle d'évolution typique pour lequel on fait varier le nombre de particules et le nombre de sites considérés, on cherche à illustrer et à quantifier ce biais.

D'une part, en approchant la log-vraisemblance par son estimation markovienne à un pas, on compare les estimations obtenues numériquement et les valeurs attendues dans les modèles descriptifs avec et sans biais. Sur l'exemple considéré, on obtient que le modèle descriptif sans biais convient dès que $m/n \leq 16$ mais qu'il n'est plus adapté dans le cas contraire. Le modèle avec biais reste en accord avec les données, même lorsque le ratio m/n est de l'ordre de 1000.

D'autre part, on exprime les deux modèles descriptifs comme des modèles de régression linéaire et on montre que le modèle sans biais ne vérifie pas la condition d'hétéroscédasticité des résidus lorsque m/n est grand alors que cette condition est vérifiée dans le modèle avec biais. On obtient aussi une quantification du biais, dont l'estimation est cohérente avec la valeur théorique énoncée dans la conjecture 8.2.6.

10.3.1 Fluctuations des estimations de $p(z_{i+1}(T) \mid z_{1:i}(T))$

Dans cette section, on vérifie pour chaque site i la convergence et la normalité asymptotique des estimateurs particulaires avec rééchantillonnage de $p(z_{i+1}(T)|z_{1:i}(T))$. De plus, on estime les écarts-types théoriques associés à chaque estimateur et on étudie les corrélations entre les estimations obtenues pour chaque site, à un ou plusieurs pas.

Données. On considère trois modèles globaux constitués à partir des modèles d'évolution, arbres et loi à la racine suivants (regroupés dans l'annexe A).

Les modèles d'évolution sont les suivants.

– $\mathsf{M}_2(1)$ est décrit par :

$$
\begin{aligned}
&v_A = 0.1576, v_C = 0.3234, v_G = 0.3380, v_T = 0.1810,\\
&w_A = 0.4959, w_C = 0.9278, w_G = 1.050, w_T = 0.4000,\\
&r_{CG\to CA} = 1.8942, r_{CA\to CG} = 0, r_{TA\to TG} = 0.3339, r_{TG\to TA} = 0.1263,\\
&r_{CA\to TA} = 0.2570, r_{CG\to TG} = 3.5230, r_{TA\to CA} = 1.9719, r_{TG\to CG} = 0.
\end{aligned}
$$

– M_5 est décrit par :

$$
\begin{aligned}
&v_A = 0.042, v_C = 0.083, v_G = 0.125, v_T = 0.167,\\
&w_A = 0.209, w_C = 0.250, w_G = 0.292, w_T = 0.334,\\
&r_{CG\to CA} = 0.104, r_{CA\to CG} = 0.730, r_{TA\to TG} = 0.438, r_{TG\to TA} = 0.730,\\
&r_{CA\to TA} = 1.501, r_{CG\to TG} = 1.835, r_{TA\to CA} = 1.627, r_{TG\to CG} = 1.877.
\end{aligned}
$$

– $\mathsf{M}_6(1)$ est décrit par :

$$\begin{aligned}
&v_A = 0.01, v_C = 0.2, v_G = 0.2, v_T = 0.01,\\
&w_A = 0.01, w_C = 0.01, w_G = 0.01, w_T = 0.01,\\
&r_{CG\to CA} = 0.2, r_{CA\to CG} = 0.2, r_{TA\to TG} = 0.2, r_{TG\to TA} = 0,\\
&r_{CA\to TA} = 0, r_{CG\to TG} = 0.2, r_{TA\to CA} = 0.2, r_{TG\to CG} = 1.
\end{aligned}$$

On choisit l'arbre $\mathsf{T}_2(0.1)$ constitué de deux arêtes de même longueur 0.1 et l'arbre T_{10} constitué de 18 arêtes et de 10 feuilles (défini précisément dans l'annexe A).

La loi à la racine R_{iid} est telle que chaque nucléotide de la séquence initiale soit indépendant des autres et selon la loi uniforme : $(0.25, 0.25, 0.25, 0.25)$.

On décrit maintenant les trois modèles globaux considérés.

Exemple 10.3.1. *(cas typique). On utilise le modèle complet* $(\mathsf{R}_{iid}, \mathsf{T}_{10}, \mathsf{M}_2(1))$. *On simule [90] selon ce modèle une séquence de feuilles, où chaque feuille est de longueur* 100.

Exemple 10.3.2. *(cas typique). On utilise le modèle* $(\mathsf{R}_{iid}, \mathsf{T}_2(0.1), \mathsf{M}_5)$. *Comme pour l'exemple 10.3.1, on simule [90] selon ce modèle une séquence de feuilles, où chaque feuille est de longueur* 100.

Exemple 10.3.3. *(cas atypique). On utilise le modèle* $(\mathsf{R}_{iid}, \mathsf{T}_2(0.1), \mathsf{M}_6(1))$ *et la séquence de feuilles de longueur* 100 *suivante :*

$$(TTT\ldots TTTY, CCC\ldots CCCY).$$

Méthode. Pour les trois exemples, on estime la quantité d'intérêt $\theta^i = p(z_{i+1}(T)|z_{1:i}(T))$ à l'aide des estimateurs particulaires $\hat{\theta}^i(n)$ avec rééchantillonnage systématique pour chaque nombre de particules n choisis dans

$$[\![2^0, 2^1, \ldots, 2^{16} = 65536]\!].$$

On répète 100 fois ces différentes estimations. Pour les 6 premiers sites de l'exemple 10.3.3, le nombre de particules maximal choisi est porté à $2^{20} = 1048576$ (avec toujours 100 répétitions).

Convergence. Pour vérifier la convergence des estimateurs particulaires du rapport de vraisemblance $p(z_{i+1}(T)|z_{1:i}(T))$, on cherche à estimer l'écart quadratique moyen entre $p(z_{i+1}(T)|z_{1:i}(T))$ et un estimateur particulaire de cette quantité. On sait qu'en général on ne dispose pas de la valeur exacte du rapport $p(z_{i+1}(T)|z_{1:i}(T))$, puisqu'il nécessite le calcul des vraisemblances $p(z_{1:i+1}(T))$ et $p(z_{1:i}(T))$, dont le coût de calcul croît exponentiellement en fonction de la longueur de la séquence (voir le corollaire 3.5.1 et la section 3.5.1). À défaut, on utilise une pseudo valeur exacte a, donnée par une estimation particulaire basée sur un nombre de particules bien supérieur à ceux pour lesquels on étudie l'écart quadratique moyen.

Définition 10.3.4. *Le pseudo écart quadratique moyen entre un estimateur* $\hat{\theta}$ *et un paramètre* θ *via la constante* a *correspond à l'écart quadratique moyen entre l'estimateur* $\hat{\theta}$ *et une estimation* a *du paramètre* θ *et est donné par :*

$$E\left((\hat{\theta} - a)^2\right).$$

Le pseudo écart quadratique de la définition 10.3.4 est ensuite estimé en utilisant un certain nombre de répétitions rép d'une estimation de la variable aléatoire $\hat{\theta}$. En notant $(\hat{\theta}_k)_{k\in[\![1,\text{rép}]\!]}$ ces estimations, on estime le pseudo écart quadratique moyen par :

$$\frac{1}{\text{rép}}\sum_{k=1}^{\text{rép}}\left(\hat{\theta}_k - a\right)^2.$$

Pour notre quantité d'intérêt $\theta^i = p(z_{i+1}(T)|z_{1:i}(T))$, on utilise les pseudo écarts suivants.

Définition 10.3.5. *Le pseudo écart quadratique $R_{partic}(n, rép)(\hat{\theta}^i)$ est le pseudo écart quadratique moyen où :*

- *$\hat{\theta}^i$ est un estimateur particulaire d'une quantité $p(z_{i+1}(T)|z_{1:i}(T))$,*
- *a est la moyenne de rép estimations particulaires indépendantes obtenues par l'estimateur particulaire avec le nombre de particules n et rééchantillonnage systématique.*

Les estimations associées au pseudo écart quadratique de la définition 10.3.5 fournissent ainsi des estimations de l'écart quadratique moyen des estimateurs particulaires de $p(z_{i+1}(T)|z_{1:i}(T))$. La pertinence de la démarche utilisée résulte de l'utilisation d'une pseudo valeur exacte a basée sur un nombre de particules bien supérieur à ceux pour lesquels on étudie l'écart quadratique moyen et du fait que l'estimateur particulaire est convergent et asymptotiquement normal lorsque le nombre de particules tend vers l'infini. Dans les simulations associées aux exemples 10.3.1, 10.3.2 et 10.3.3, on choisit pour estimer la pseudo valeur exacte a un nombre de particules égal à 2^{16} ou 2^{20} et 100 répétitions.

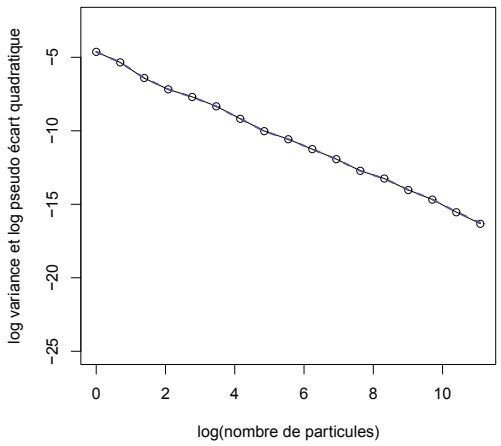
Pour les exemples 10.3.1 et 10.3.2, pour chaque site i et chaque estimateur $\hat{\theta}^i(n)$ associé au nombre de particules n, on calcule $R_{\text{partic}}(2^{16}, 100)(\hat{\theta}^i(n))$ le pseudo écart quadratique moyen ainsi que les variances empiriques. Pour l'exemple 10.3.1, on représente sur la figure 10.6 l'évolution de ces quantités en fonction du nombre de particules pour quatre sites.

Pour l'exemple 10.3.3, pour les six premiers sites et chaque estimateur $\hat{\theta}^i(n)$ associé au nombre de particules n, on calcule le pseudo écart quadratique moyen $R_{\text{partic}}(2^{20}, 100)(\hat{\theta}^i)$ ainsi que les variances empiriques. On représente sur la figure 10.7 l'évolution de ces quantités en fonction du nombre de particules pour les sites 2, 3, 4 et 6.

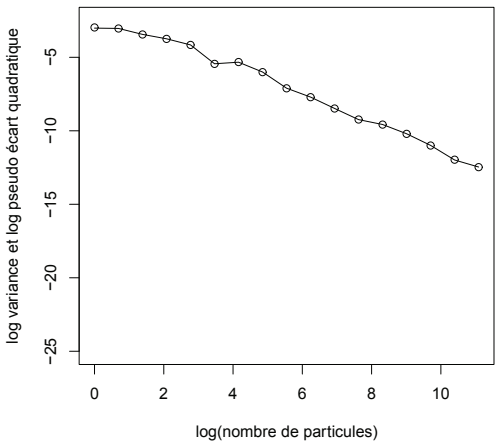
Pour le site 2, le calcul explicite du rapport de vraisemblance $\theta^i = p(z_{i+1}(T)|z_{1:i}(T))$ (avec $i = 1$) est accessible et effectué (les calculs matriciels faisant intervenir des exponentielles de matrices 36×36, voir section 3.5.1). On en déduit pour ce site l'écart quadratique moyen entre les estimateurs $\hat{\theta}^i(n)$ et la valeur exacte θ^i. On représente sur les figures 10.6 et 10.7, pour le site 2, l'évolution de l'écart quadratique moyen en fonction du nombre de particules.

Dans les trois exemples, on observe pour le site 2 que le comportement du pseudo écart quadratique $R_{\text{partic}}(\hat{\theta}^i(n))$ est semblable (dès $n = 2^3$ particules) à celui de l'écart quadratique, et également semblable à la variance empirique. Ainsi, le biais entre les estimateurs $\hat{\theta}^i(n)$ et le paramètre d'intérêt θ^i n'est pas perceptible pour ce site.

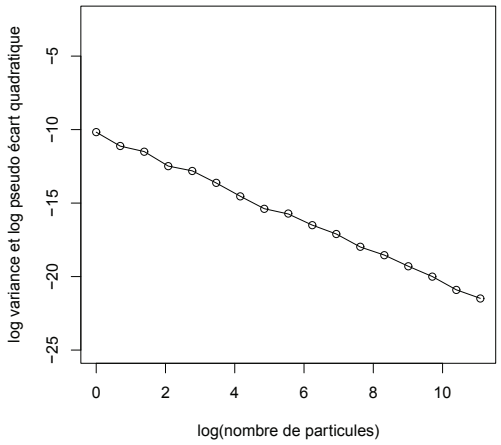
Pour tous les sites, on observe dans les trois exemples que le comportement du pseudo écart quadratique $R_{\text{partic}}(\hat{\theta}^i(n))$ en fonction du nombre de particules n est semblable à celui de la variance empirique, avec une décroissance en $1/n$ comme attendue.



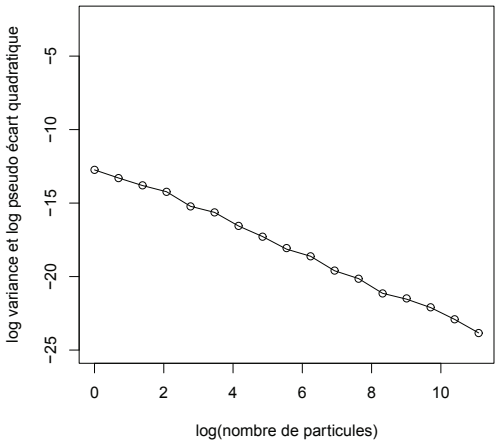
Site 2.



Site 19.

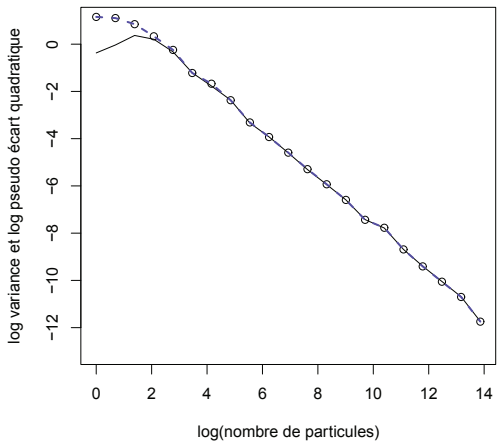


Site 28.

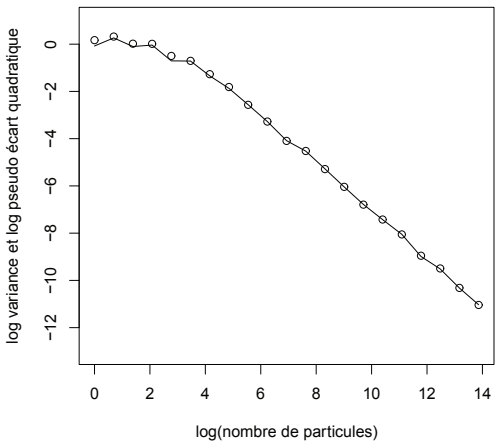


Site 77.

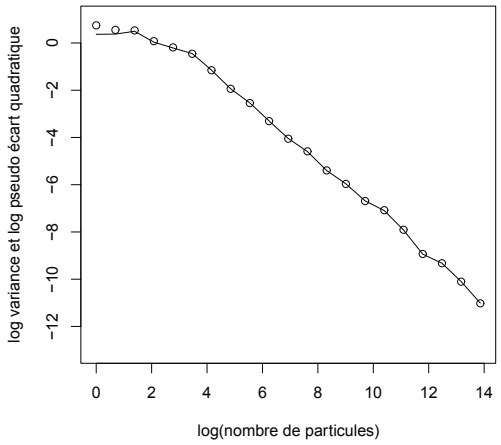
FIGURE 10.6 – Évolution pour quatre sites du pseudo écart quadratique moyen R_{partic} (points) et de la variance empirique (ligne noire) en fonction du nombre de particules, sur un repère log-log, associé à l'exemple 10.3.1. Pour le graphique associé au site 2 est aussi représenté l'écart quadratique moyen (pointillés bleus).



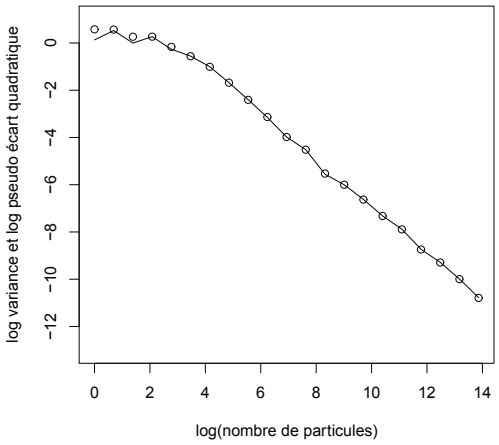
Site 2.



Site 3.



Site 4.



Site 6.

FIGURE 10.7 – Évolution pour quatre sites du pseudo écart quadratique moyen R_{partic} (points) et de la variance empirique (ligne noire) en fonction du nombre de particules, sur un repère log-log, associé à l'exemple 10.3.3. Pour le graphique associé au site 2 est aussi représenté l'écart quadratique moyen (pointillés bleus).

Estimation des écarts-types. On déduit les différents écarts-types empiriques à partir du calcul des variances empiriques. On illustre la décroissance en $1/\sqrt{n}$ des fluctuations telle qu'énoncée par le théorème 8.1.8 par la figure 10.8, en représentant sur un repère log-log pour deux sites de l'exemple 10.3.1 l'écart-type empirique en fonction du nombre de particules utilisées, ainsi que des droites de régression associées.

Une estimation des écarts-types $\sigma^i_{\text{APF optimal}}$ théoriques du théorème 8.1.8 en chaque site i peut être envisagée. Pour chaque site, on effectue pour cela une régression linéaire avec pente $-1/2$ du logarithme de l'écart-type en fonction du logarithme du nombre de particules. Pour l'exemple 10.3.1, les écarts-types estimés sont représentés sur la figure 10.9.

Pour les sites 62, 73, 79 et 83, toutes les feuilles sont égales au dinucléotide encodé RY. L'estimation de l'écart-type devrait alors être -∞, mais reste proche de -14.5 dû au nombre de chiffres significatifs limité du type `double` en C++.

Notons que la méthode proposée pour estimer $\sigma^i_{\text{APF optimal}}$ repose uniquement sur l'échantillon d'écarts-types empiriques obtenu, sans chercher à exploiter les expressions explicites de la variance fournis dans le théorème 8.1.8. Ce choix est dû d'une part par la facilité de mise en œuvre de la méthode proposée et d'autre part par la difficulté d'estimer les quantités $p(x_{1:k}|y_{1:i}) = p(z_{1:k}|z_{1:i}(T))$ de l'équation (8.1) pour $k \in [\![1, i-1]\!]$ (par exemple pour la quantité $p(z_1|z_{1:i}(T))$).

Normalité. Pour chacun des trois exemples, pour chaque choix du nombre de particules et chaque site, on calcule la p-value pour le test de normalité de Shapiro-Wilk [104] sur l'échantillon de taille 100 (correspondant aux 100 répétitions indépendantes). On compare ensuite chaque valeur obtenue avec cinq pour cent. On représente sur les figures 10.10 et 10.11 la proportion de sites avec une p-value plus grande que 0.05 en fonction du nombre de particules choisis pour les exemples 10.3.1 et 10.3.3 (la figure correspond à l'exemple 10.3.2 est semblable à la figure 10.10 et n'est pas représentée ici).

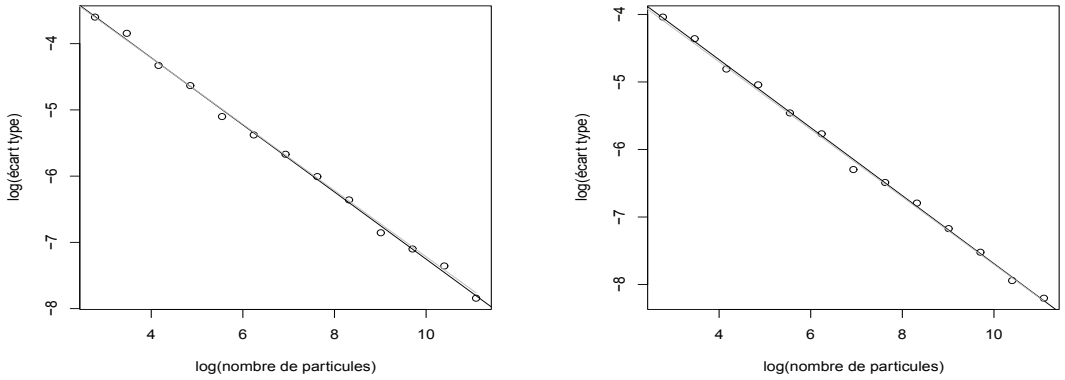
On sait que d'après le théorème 8.1.8 que s'il y avait indépendance entre les différents sites, on s'attendrait à ce que la proportion se fixe autour de 0.95 quand le nombre de particules n tend vers l'infini. Bien qu'il n'y ait en général pas indépendance entre les sites de la séquence, on observe dans les trois exemples que cette proportion augmente et devient proche de 0.95. On en conclut que l'on ne peut pas rejeter l'hypothèse de normalité avec un risque 0.05.

Dans l'exemple 10.3.3 atypique, on constate dans ce contexte de forte dépendance un profil différent de celui obtenu dans les exemples 10.3.1 et 10.3.2.

Corrélations entre les échantillons de chaque site.

Corrélation à un pas. On reprend les exemples 10.3.1, 10.3.2 et 10.3.3 et le choix de 65536 particules. On dispose pour chaque site de 100 estimations de $\log p(z_{i+1}(T)|z_{1:i}(T))$.

On déduit pour chaque site $i \in [\![1, m-1]\!]$ la corrélation empirique à un pas c_i des estimations du site i avec celles du site $i+1$. On pose et on calcule $c = (c_i)_{i \in [\![1, m-2]\!]}$, puis on en déduit la moyenne empirique $\bar{c}$, l'écart-type empirique $\sigma(c)$ et le maximum en valeur absolue de la suite de valeur c. On regroupe les valeurs obtenues dans le tableau suivant :



(a) Site 13. (b) Site 14.

FIGURE 10.8 – Évolution du logarithme de l'écart-type empirique de l'exemple 10.3.1 en fonction du logarithme du nombre de particules utilisés, pour deux sites (points). La ligne noire correspond à la régression linéaire de ces points et la ligne grise à la régression linéaire avec une pente fixée à $-1/2$.

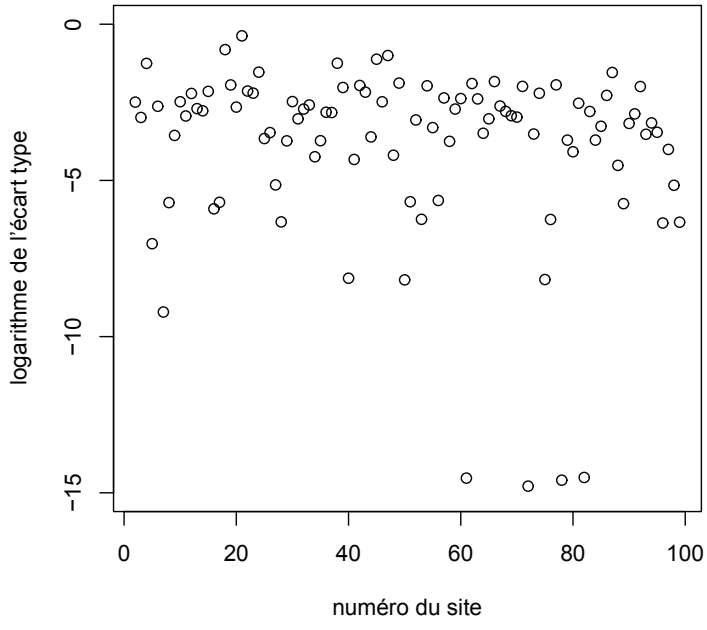


FIGURE 10.9 – Estimation pour l'exemple 10.3.1 du logarithme de l'écart-type $\sigma^i_{\text{APF optimal}}$ en chaque site.

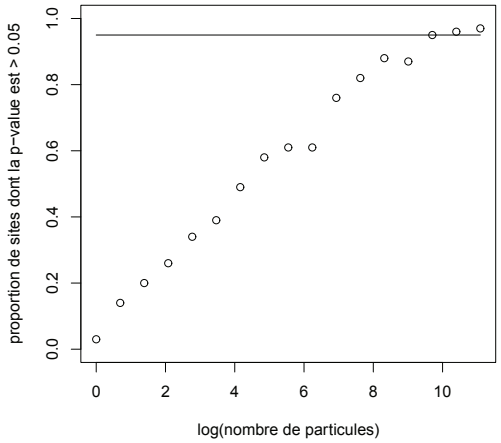


FIGURE 10.10 – Exemple 10.3.1.

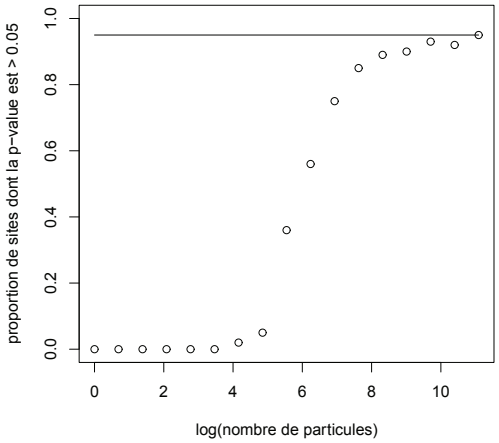


FIGURE 10.11 – Exemple 10.3.3.

Proportion de sites avec une p-value supérieure à 0.05 dans le test de normalité de Shapiro-Wilk, en fonction du logarithme du nombre de particules.

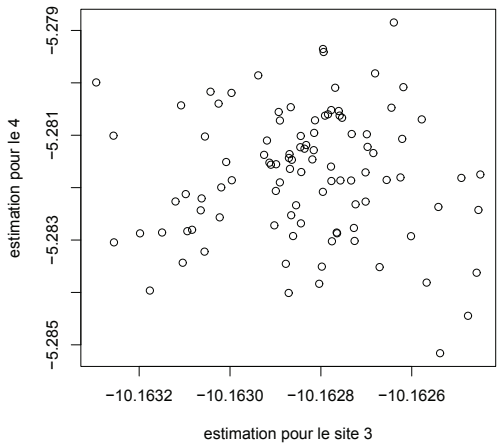
	c_4	$\bar{c}$	$\sigma(c)$	$\max \lvert c \rvert$
exemple 10.3.1 (typique)	0.109	-0.001	0.092	0.211
exemple 10.3.2 (typique)	-0.098	-0.003	0.105	0.296
exemple 10.3.3 (atypique)	-0.556	-0.539	0.079	0.683

Pour les exemples 10.3.1 et 10.3.3, on observe sur la figure 10.12 les estimations de $\log p(z_4(T)|z_{1:3}(T))$ en fonction de $\log p(z_3(T)|z_{1:2}(T))$.

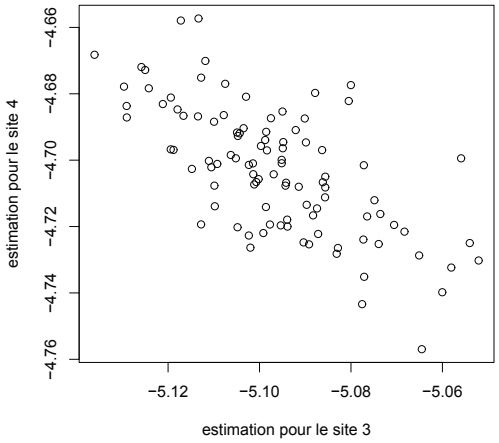
On observe que dans les exemples 10.3.1 et 10.3.2, aucune corrélation linéaire entre deux sites consécutifs ne semble apparaître alors qu'elle semble importante dans l'exemple 10.3.3. Pour confirmer cela, on effectue entre chaque couple de sites consécutifs $(i, i+1)$ le test de significativité de corrélation de Pearson. On en déduit la p-value associée à chacun de ces tests et on calcule la proportion de couples dont la p-value associée est supérieure à 0.05. Dans le cas où il n'y a pas de corrélation linéaire visible, on s'attend à obtenir une proportion de couples proche de 0.95. Pour les trois exemples, on obtient les valeurs 0.98 (exemple 10.3.1), 0.97 (exemple 10.3.2) et 0.01 (exemple 10.3.3). Cela confirme que la corrélation entre deux sites consécutifs est significative uniquement dans l'exemple atypique 10.3.3.

Notons que pour l'exemple atypique 10.3.3, la corrélation est toujours négative. Cela a pour conséquence de recentrer la somme de deux valeurs de log-vraisemblance consécutives autour de sa valeur moyenne. Notons également que globalement, on observe les mêmes conclusions pour chaque choix du nombre de particules supérieur à $2^3 = 8$.

Corrélation à portée longue. De la même manière que pour l'étude des corrélations à un pas, on effectue pour $k \in [\![1, 10]\!]$ le test de significativité de corrélation de Pearson entre les sites i et $i+k$. On en déduit la p-value associée à chacun de ces tests et on calcule



Exemple 10.3.1.



Exemple 10.3.3.

FIGURE 10.12 – Estimations de $\log p(z_4(T)|z_{1:3}(T))$ en fonction de $\log p(z_3(T)|z_{1:2}(T))$.

la proportion de couples dont la p-value associée est supérieure à 0.05. On regroupe dans le tableau suivant les différentes valeurs obtenues. On observe que même dans l'exemple atypique 10.3.3, les corrélations entre deux sites à distance k deviennent non significatives lorsque le pas k devient grand (pour $k \geq 8$ dans l'exemple 10.3.3).

k	1	2	3	4	5	6	7	8	9	10
exemple 10.3.1 (typique)	0.98	0.94	0.95	1.00	0.97	0.91	0.92	0.92	0.98	0.95
exemple 10.3.2 (typique)	0.97	0.94	0.91	0.96	0.95	0.94	0.91	0.94	0.98	0.95
exemple 10.3.3 (atypique)	0.01	0.10	0.56	0.84	0.87	0.88	0.90	0.96	0.96	0.94

10.3.2 Fluctuations des estimations de vraisemblance de la séquence

On souhaite non plus regarder le comportement asymptotique d'un terme $\log p(z_{i+1}(T) \mid z_{1:i}(T))$ mais le comportement de la somme de ces quantités, c'est-à-dire l'estimation globale de la vraisemblance $\log p(z_{1:m}(T))$. Sur les exemples 10.3.1 et 10.3.3 déjà utilisés dans la section 10.3.1, on regarde la convergence et la normalité asymptotique des estimateurs particulaires de $\log p(z_{1:m}(T))$ avec rééchantillonnage systématique.

Données. On reprend les exemples 10.3.1 et 10.3.3 de la section 10.3.1, et on reprend les données simulées. Pour chaque choix de particules n, on déduit 100 estimations particulaires associées à l'estimateur particulaire de la log-vraisemblance $\hat{\theta}(n) := \hat{L}_{n,\text{1-partic}}$ à partir des quantités issues des estimateurs $\hat{\theta}^i(n)$ en écrivant :

$$\hat{\theta}(n) = \sum_{i=0}^{m-1} \log \hat{\theta}^i(n).$$

Méthode. On raisonne de manière analogue à ce qui a été fait dans la section 10.3.1. On veut estimer l'écart quadratique moyen entre la log-vraisemblance $\log p(z_{1:m}(T))$ et un estimateur de cette quantité. Comme on ne dispose pas en général de la valeur exacte de cette log-vraisemblance, on utilise une pseudo valeur exacte a donnée par une estimation particulaire basée sur un grand nombre de particules. La démarche reste pertinente puisque l'estimateur particulaire est convergent et asymptotiquement normal (voir théorème 8.2.5) lorsque le nombre de particules tend vers l'infini.

En adaptant la définition 10.3.5 de la section 10.3.1, on définit le pseudo écart quadratique pour la quantité d'intérêt $\theta = \log p(z_{1:m}(T))$:

Définition 10.3.6. *Le pseudo écart quadratique $R^{1:m}_{partic}(n, rép)(\hat{\theta})$ est le pseudo écart quadratique moyen où :*

- *$\hat{\theta}$ est un estimateur particulaire d'une quantité $\log p(z_{1:m}(T))$,*
- *a est la moyenne de rép estimations particulaires indépendantes obtenues par l'estimateur particulaire avec le nombre de particules n et rééchantillonnage systématique.*

Convergence, normalité asymptotique et estimation de la variance.

On calcule le pseudo écart quadratique $R^{1:m}_{\text{partic}}(n, \text{rép})(\hat{\theta})$ ainsi que les variances empiriques des estimations particulaires de $\log p(z_{1:m}(T))$ pour chaque choix du nombre de particules n. On représente sur la figure 10.13 l'évolution de ces quantités en fonction du nombre de particules.

On observe sur cette figure un comportement cohérent des estimateurs particulaires de $\log p(z_{1:m}(T))$ avec rééchantillonnage systématique dans les deux exemples, avec dès $2^{10} = 1024$ particules utilisées une décroissance du pseudo écart quadratique $R^{1:m}_{\text{partic}}$ en $1/n$. De plus, on n'observe pas de biais apparent graphiquement (entre l'estimateur $\hat{\theta}(n)$ et la pseudo valeur exacte a). Cela traduit, avec ε une borne entre le logarithme du pseudo écart quadratique moyen et le logarithme de la variance :

$$\log E((\theta - a)^2) - \log \operatorname{Var} \hat{\theta} \leq \varepsilon$$

et en écrivant :

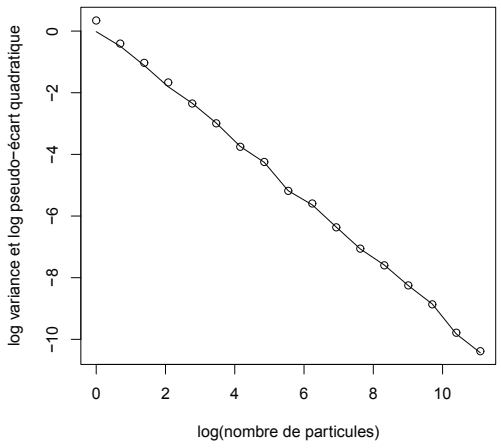
$$E((\hat{\theta} - a)^2) = \operatorname{Var} \hat{\theta} + (E\hat{\theta} - a)^2,$$

que le biais entre $\hat{\theta}$ et la valeur a est borné de la façon suivante :

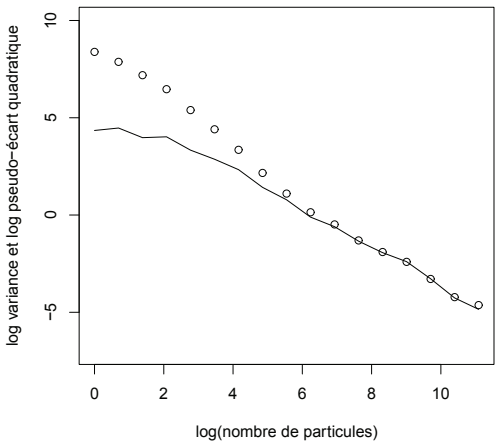
$$E\hat{\theta} - a \leq \sqrt{e^{\varepsilon} - 1} \sqrt{\operatorname{Var} \hat{\theta}}.$$

On dispose pour chaque exemple de 100 estimations de la log-vraisemblance pour chaque choix du nombre de particules. Le test de Shapiro-Wilk montre que dans les deux exemples et pour tous les choix de particules, on ne peut pas rejeter l'hypothèse de normalité avec un risque 0.01. Cela est cohérent avec le théorème 8.2.5.

De la même manière que dans la section 10.3.1, on déduit des échantillons une estimation de la variance théorique (c'est-à-dire multipliée par $\sqrt{n}$ où n est un choix du nombre de particules) de l'estimateur particulaire de $\log p(z_{1:m}(T))$ (voir théorème 8.2.5). De nouveau, cette estimation résulte directement de l'échantillon et on ne cherche pas à



Exemple 10.3.1.



Exemple 10.3.3.

FIGURE 10.13 – Évolution du pseudo écart quadratique moyen $R^{1:m}_{\text{partic}}$ (points) et de la variance empirique des estimations particulaires de $\log p(z_{1:m}(T))$ (ligne noire) en fonction du nombre de particules et sur un repère log-log.

utiliser une formule théorique spécifique aux méthodes particulaires de la variance. Pour l'exemple typique 10.3.1, on obtient 1.68. Comme attendu d'après l'étude des corrélations entre les sites, cette valeur est du même ordre de grandeur que la somme des estimations de variance de l'estimateur particulaire $p(z_{i+1}(T)|z_{1:i}(T))$ en chaque site, qui est de 1.54.

Pour l'exemple atypique 10.3.3, on obtient une variance de 37.11, sensiblement plus petite que la somme 64.20 des variances estimées en chaque site. D'après l'étude des corrélations entre les sites, cet écart semble résulter de la corrélation négative entre chaque paire de sites consécutifs.

10.3.3 Présence et estimation du biais

On cherche dans cette section à illustrer le biais dans l'estimation de la log-vraisemblance en fonction du nombre de sites m et de particules n considérés. On veut de plus vérifier ou infirmer les formes du biais attendus décrits par le théorème 8.2.5 (dans le cas où n est grand devant m) et la conjecture 8.2.6 (dans le cas où le nombre de particules n est du même ordre de grandeur ou plus petit que le nombre de sites m).

On considère alors les données et les estimations particulaires issues de l'exemple suivant :

Exemple 10.3.7. *On choisit le modèle d'évolution* $\mathsf{M}_2(1)$ *typique déjà défini (voir l'annexe A), la loi à la racine* R_{iid} *(telle que chaque nucléotide de la séquence initiale soit indépendant des autres et selon la loi uniforme) et l'arbre* T_6 *constitué de* 6 *feuilles et défini avec la notation de Newick [5] comme :*

$\mathsf{T}_6 = (Bovine : 0.69395, (Hylobates : 0.36079, (Pongo : 0.33636, (GGorilla : 0.17147, (Ppaniscus : 0.19268, Hsapiens : 0.11927) : 0.08386) : 0.06124) : 0.15057) : 0.54939);$

Le modèle complet s'écrit $(\mathsf{R}_{iid}, \mathsf{T}_6, \mathsf{M}_2(1))$ *et on simule une séquence de longueur* $2^{11} = 2048$ *obtenue par simulation [90].*
Pour chaque choix d'un nombre de sites $m \in [\![2^1, \ldots, 2^{11}]\!]$ *et de particules* $n \in [\![2^0, \ldots, 2^{13}]\!]$, *on répète* 100 *fois l'estimation de la log-vraisemblance des* m *premiers sites avec* n *particules et rééchantillonnage systématique. On note* $\hat{L}(m, n)$ *la moyenne de ces estimations.*

On considère également $\hat{L}_{Markov}(m)$ *l'estimation de la log-vraisemblance des* m *premiers sites par approximation markovienne à* 1 *pas.*

On suppose dans toute cette section que le nombre de sites m est fixé. On note $L(m)$ la valeur exacte de la log-vraisemblance et γ une variable aléatoire de loi normale centrée réduite. On cherche à exprimer les estimateurs particulaires $\hat{L}(m, n)$ (où n est le nombre de particules) en fonction de la valeur exacte $L(m)$. On note également σ l'écart-type théorique associé aux estimateurs $\hat{L}(m, n)$. On considère alors deux modèles. Le premier est un modèle sans biais, adapté de l'énoncé du théorème 8.2.5, défini par l'équation suivante :

$$\hat{L}(m, n) = L(m) + \sigma\sqrt{\frac{m}{n}}\gamma. \tag{10.5}$$

Le deuxième est un modèle avec prise en compte d'un biais $b \in \mathbb{R}$ proportionnel à m/n, défini de la façon suivante :

$$\hat{L}(m, n) = L(m) + b\frac{m}{n} + \sigma\sqrt{\frac{m}{n}}\gamma. \tag{10.6}$$

Le choix proposé pour b par la conjecture 8.2.6 est de considérer $b = -\sigma^2/2$.

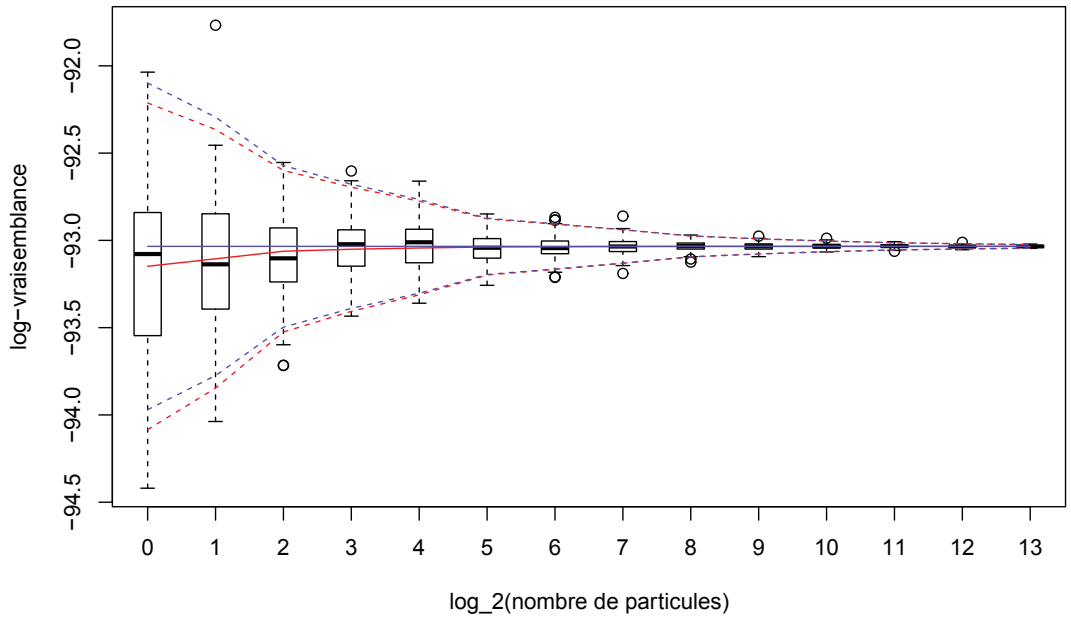
Utilisation de l'estimation par approximation markovienne comme valeur de référence. On fait l'hypothèse que l'on peut négliger l'écart entre la vraie valeur de log-vraisemblance et l'estimation obtenue par l'approximation markovienne $\hat{L}_{\text{Markov}}(m)$. De plus, on utilise ici $b = -\sigma^2/2$ comme choix du biais. On cherche alors à comparer visuellement les deux modèles suivants :

$$\hat{L}(m, n) = \hat{L}_{\text{Markov}}(m) + \sigma\sqrt{\frac{m}{n}}\gamma, \tag{10.7}$$

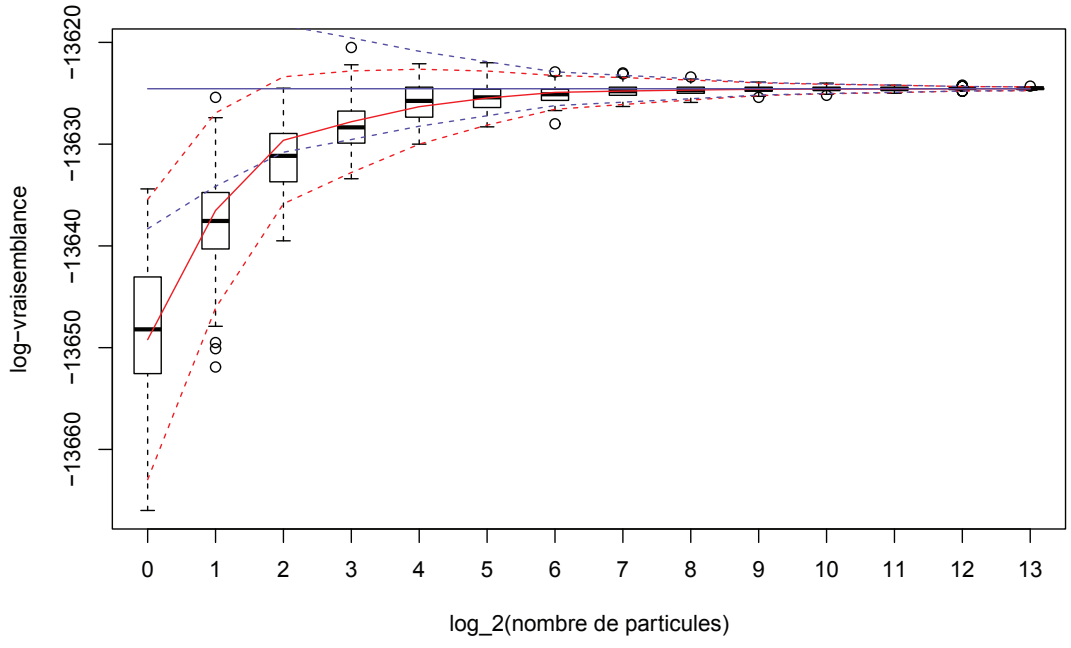
$$\hat{L}(m, n) = \hat{L}_{\text{Markov}}(m) - \frac{\sigma^2}{2}\frac{m}{n} + \sigma\sqrt{\frac{m}{n}}\gamma. \tag{10.8}$$

Pour deux choix du nombre de sites ($m = 16$ et $m = 2048$), pour chaque choix du nombre de particules n, on considère les 100 estimations de la log-vraisemblance obtenus dans l'exemple 10.3.7. D'une part, on trace sur la figure 10.14 le diagramme en boîte associé et d'autre part, on estime empiriquement la valeur σ. On représente ensuite les courbes espérées et les intervalles de confiance associés aux deux modèles 10.7 et 10.8 (en bleu pour le modèle sans biais et en rouge pour le modèle avec biais).

Pour cet exemple, on observe dans les deux cas que les deux modèles 10.7 et 10.8 concordent avec les données dès $n = 2^7 = 128$ particules. Par contre lorsque $m/n > 16$, le modèle sans prise en compte du biais n'est plus en accord avec les données alors que le modèle avec prise en compte du biais semble toujours valide.



$m = 16$.



$m = 2048$.

FIGURE 10.14 – Pour deux choix du nombre de sites dans l'exemple de la section 10.3.3, diagrammes en boîtes associés associés aux 100 estimations de la log-vraisemblance en fonction du nombre de particules utilisées. La ligne bleue (resp. la ligne rouge) représente les différentes valeurs espérées dans le modèle de l'équation (10.7) sans prise en compte du biais (resp. dans le modèle de l'équation (10.8) avec prise en compte du biais). Les lignes pointillées représentent les intervalles de confiance à 95% associés.

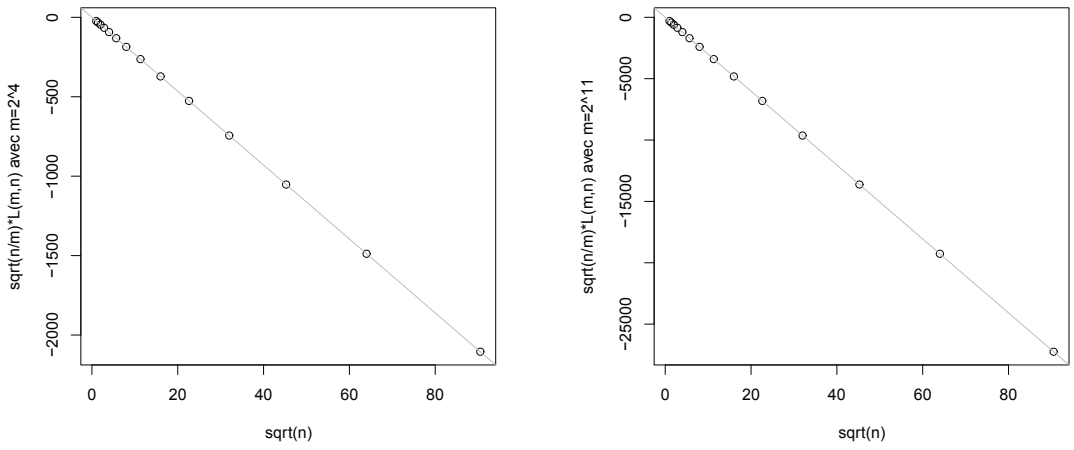
Modèle de régression sans biais. On reprend le modèle initial sans biais (10.5) :

$$\hat{L}(m,n) = L(m) + \sigma\sqrt{\frac{m}{n}}\gamma.$$

On multiplie l'équation par $\sqrt{\frac{n}{m}}$ et pour chaque m fixé, on considère le modèle suivant de régression linéaire (avec $\sqrt{n}$ la variable explicative) :

$$\sqrt{\frac{n}{m}}\hat{L}(m,n) = \sqrt{n}\left(\frac{L(m)}{\sqrt{m}}\right) + \sigma\gamma.$$

Après avoir effectué pour chaque m les régressions de $\sqrt{\frac{n}{m}}\hat{L}(m,n)$ en fonction de $\sqrt{n}$ en imposant une ordonnée à l'origine nulle, on obtient que tous les paramètres de pente à estimer vérifient les tests de significativité. On observe que les écarts obtenus entre l'estimation de $L(m)$ et l'estimation markovienne considérée crédible $\hat{L}_{\text{Markov}}(m)$ sont tous inférieurs à 0.1 (par exemple, pour $m = 2^{11}$, on obtient $L(m) = -13624.54$ et $\hat{L}_{\text{Markov}}(m) = -13624.55$). On représente sur la figure 10.15 les droites de régression obtenues pour deux choix du nombre de sites.



$m = 16$. $m = 2048$.

FIGURE 10.15 – Nuage de points et droite de régression associée (ligne grise) pour le modèle sans prise en compte du biais de la section 10.3.3, pour deux choix de longueur de la séquence.

Par contre, les résidus ne vérifient pas la condition d'hétéroscédasticité classique lorsque le nombre de sites devient du même ordre de grandeur que le nombre de particules considérés comme le montre la figure 10.16. On introduit alors un nouveau modèle tenant compte du biais présent quand le nombre de sites est du même ordre de grandeur ou plus grand que le nombre de particules.

Modèle de régression avec biais. On reprend le modèle initial avec biais (10.5) :

$$\hat{L}(m,n) = L(m) + b\frac{m}{n} + \sigma\sqrt{\frac{m}{n}}\gamma.$$

En multipliant une nouvelle fois par $\sqrt{\frac{n}{m}}$, on considère pour chaque m fixé le modèle de régression linéaire suivant (avec $\sqrt{n}$ et $\frac{1}{\sqrt{n}}$ les variables explicatives) :

$$\sqrt{\frac{n}{m}}\hat{L}(m,n) = \sqrt{n}\left(\frac{L(m)}{\sqrt{m}}\right) + \frac{1}{\sqrt{n}}\left(b\sqrt{m}\right) + \sigma\gamma.$$

Pour chaque m fixé, on effectue l'estimation des paramètres associés à cette régression linéaire, qui vérifient tous les tests de significativité. De plus, le problème de comportement hétéroscédastique des résidus est éliminé (voir figure 10.16). On note pour chaque choix de m la valeur b_m estimée du biais b. Pour un nombre de sites m supérieur à 64, la constante b est estimée par environ -0.01.

Pour chaque m, on compare b_m avec $-v_m/2$, où v_m est la variance estimée de la log-vraisemblance associée à l'échantillon de taille 100 obtenu avec $n = m$. On représente les points associés sur la figure 10.17. On observe que les estimations obtenues par régression linéaire sont cohérentes avec la présence d'un biais donné par $-v/2$, avec v la variance de la log-vraisemblance.

10.4 Inférence d'un nucléotide de la racine

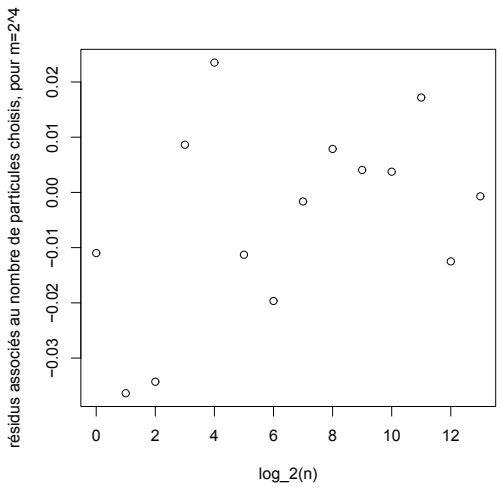
Le but de cette section correspond à une problématique différente par rapport aux sections précédentes puisque l'on cherche, pour un modèle global fixé, à inférer la séquence ancestrale à partir des observations.

Plusieurs approches sont possibles pour rechercher la *meilleure* séquence ancestrale. On peut par exemple inférer pour chaque site le nucléotide ancestral le plus probable et reconstituer ensuite une séquence ancestrale, ou bien trouver directement la séquence la plus probable globalement au vu des données (voir aussi le paragraphe *B. Solution to Problem 2* dans [99]).

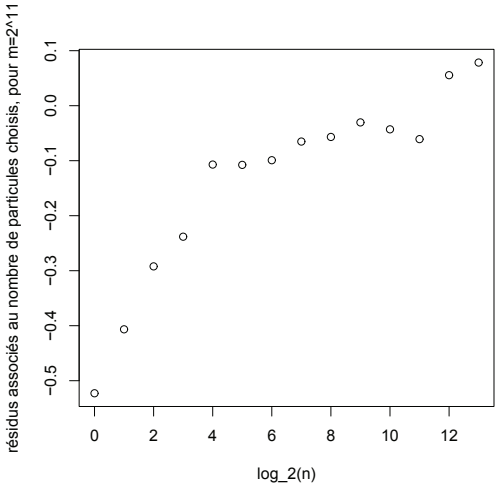
Dans cette section, on raisonne site par site en inférant pour chaque site le nucléotide ancestral le plus probable vis-à-vis des observations. Deux méthodes sont proposées dans la section 10.4.1, toutes les deux basées sur les évolutions particulaires. Elles sont ensuite comparées dans la section 10.4.2. On observe que les deux méthodes sont convergentes et peuvent être utilisées pour estimer la loi du nucléotide le plus probable en un site. Comme la deuxième méthode s'applique de façon plus générale, elle est préférée par la suite.

L'utilisation des méthodes particulaires étant coûteux, on cherche à prendre en compte le voisinage relativement proche en négligeant ce qui se trouve assez loin dans le but de se ramener à des séquences de longueur 3 et 5 pour lesquels le calcul matriciel exact est numériquement possible. Pour cela, on cherche à quantifier dans les sections 10.4.3 et 10.4.4 le nombre de sites voisins à prendre en compte autour du site considéré pour inférer de façon pertinente le nucléotide ancestral en un site.

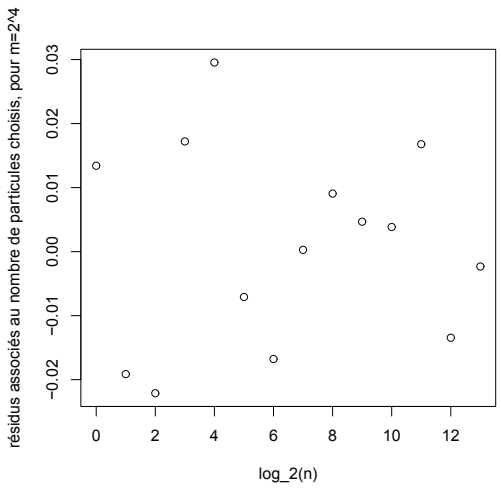
Dans la section 10.4.3, on utilise un modèle atypique qui a été obtenu par une étude exploratoire des paramètres du modèle. On met alors en évidence un modèle associé à des séquences atypiques qui nécessitent de considérer plus d'une dizaine de nucléotides pour ne pas faire d'inférence erronée. Cela permet de montrer que l'on ne peut pas se contenter de négliger les observations au-delà de deux pas autour du site considéré. Cela montre



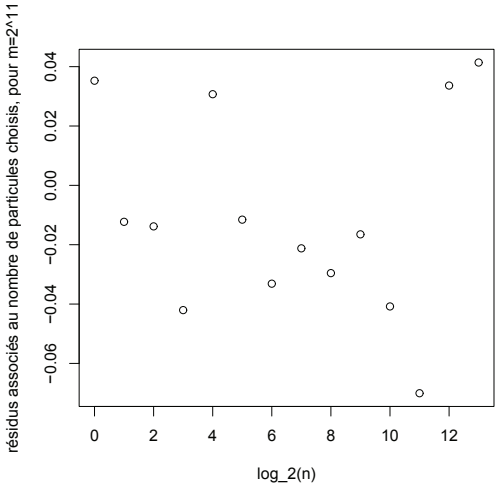
Modèle sans prise en compte du biais, $m = 16$.



Modèle sans prise en compte du biais, $m = 2048$.



Modèle avec prise en compte du biais, $m = 16$.



Modèle avec prise en compte du biais, $m = 2048$.

FIGURE 10.16 – Visualisation des résidus pour les modèles sans et avec prise en compte du biais de la section 10.3.3, pour deux choix de longueur de la séquence.

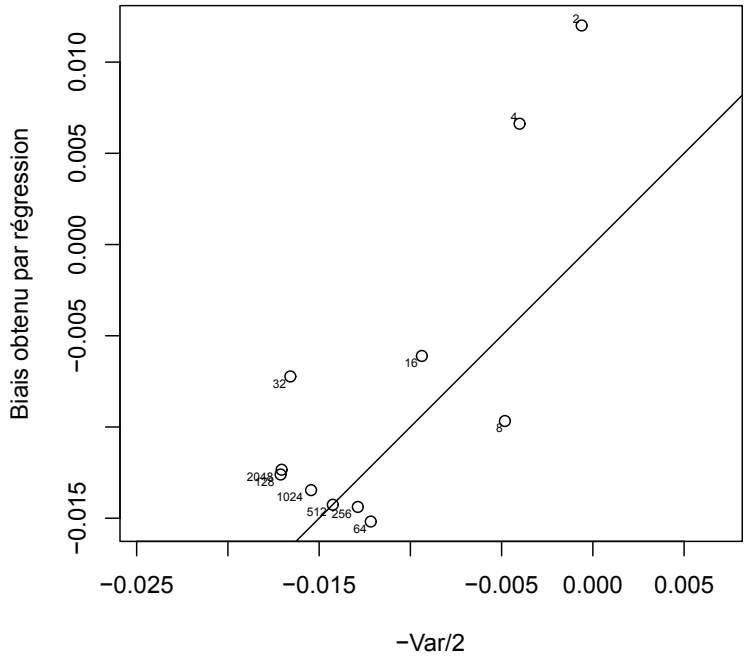


FIGURE 10.17 – Pour chaque choix du nombre de sites m, on place le point d'abscisse $-v_m/2$ et d'ordonnée b_m. On trace de plus la droite d'équation $y = x$.

également que les évolutions particulaires utilisées permettent ici d'obtenir des estimations précises et en temps raisonnable des différentes probabilités recherchées.

Dans la section 10.4.4, on utilise un modèle typique et on montre de façon générale que l'influence des voisins au-delà de deux pas peut être négligée. Dans ce cas, on peut éviter d'utiliser des méthodes particulaires et effectuer uniquement le calcul matriciel associé aux séquences de longueur 5 autour du site considéré.

Enfin, dans la section 10.4.5, on propose une approche d'inférence d'un nucléotide de la racine permettant de prendre en compte le voisinage relativement proche en négligeant ce qui se trouve en dehors. Cette approche consiste à utiliser par défaut un nombre de pas restreint pour inférer le nucléotide voulu, et d'utiliser les évolutions particulaires lorsqu'une incertitude dans l'inférence apparaît. Un algorithme pour reconstituer la séquence ancestrale complète est ensuite proposé.

10.4.1 Méthodes d'inférence à la racine

On choisit un modèle complet $\lambda = (\mathsf{R}, \mathsf{T}, \mathsf{M})$ (voir section 1.3), avec M un modèle RN95+YpR, ainsi qu'un jeu de séquences observées $\mathbf{y}$ associées aux feuilles de l'arbre.

On suppose que la racine R est la loi stationnaire du modèle M, une approximation markovienne de cette loi stationnaire ou une loi définie indépendamment sur chaque site. Dans les trois cas, la loi à la racine R vérifie l'hypothèse 6.4.1 énoncée dans la section 6.4.1, c'est-

à-dire qu'il est possible de découpler pour chaque site l'évolution à la racine et l'évolution conditionnellement à la racine.

Le but de cette section est de définir deux méthodes permettant d'approcher conditionnellement aux observations la loi du nucléotide à la racine en un site $i \in [\![1, m]\!]$ fixé, c'est-à-dire à calculer pour $N \in \mathcal{A}$ (avec $X_i(0)$ la variable du nucléotide au site i et au temps 0) :

$$P_\lambda(X_i(0) = N \mid \mathbf{y}).$$

Ces deux méthodes sont basées sur la possibilité de simuler des évolutions particulaires selon un certain modèle d'évolution RN95+YpR et conditionnellement aux observations (voir section 8.2).

Méthode 1. Pour utiliser cette méthode, pour chaque $N \in \mathcal{A}$ on définit R_N le modèle à la racine conditionné à valoir N au site i. On pose $\lambda' = (\mathsf{R}_N, \mathsf{T}, \mathsf{M})$. En utilisant l'hypothèse 6.4.1, on en déduit que le modèle λ' vérifie la proposition 6.4.3 de découplage.

On fait l'hypothèse supplémentaire que l'on est capable de connaître la loi de R_N, ce qui est immédiat uniquement dans le cas où la loi à la racine est définie indépendamment sur chaque site. On écrit dans ce cas :

$$P_\lambda(X_i(0) = N \mid \mathbf{y}) = \frac{P_\lambda(\mathbf{y} \mid X_i(0) = N) P_\lambda(X_i(0) = N)}{\sum_{N'} P_\lambda(\mathbf{y} \mid X_i(0) = N') P_\lambda(X_i(0) = N')}. \tag{10.9}$$

Comme la loi de R_N est connue, on peut calculer avec les estimateurs particulaires les probabilités $P_\lambda(\mathbf{y} \mid X_i(0) = N) = P_{\lambda'}(\mathbf{y})$, pour tout $N \in \mathcal{A}$. On en déduit ensuite pour tout $N \in \mathcal{A}$ les probabilités recherchées $P_\lambda(X_i(0) = N \mid \mathbf{y})$.

Avec cette méthode, on remarque que l'on doit effectuer quatre évolutions particulaires (une pour chaque nucléotide) pour obtenir une approximation de la loi du nucléotide à la racine en un site i.

Méthode 2. Dans cette méthode, on effectue tout d'abord des évolutions particulaires sous le modèle λ conditionnellement aux observations $\mathbf{y}$. En particulier, le modèle à la racine considéré n'est pas modifié.

En choisissant une évolution particulaire à n particules, on dispose d'un échantillon $(z^j_{1:m-1})_{j \in [\![1,n]\!]}$ d'évolutions issues de $Z_{1:m-1}$ conditionnées aux observations. On va alors regarder pour chaque particule j si la racine $x^j_i(0)$ au site i associée à l'évolution $(z^j_{1:m-1})$ est N. On en déduit la proportion de particules telles que la racine au site i soit le nucléotide N, c'est-à-dire :

$$\frac{1}{n} \sum_{j=1}^{n} \mathbf{1}(x^j_i(0) = N). \tag{10.10}$$

Par la loi des grands nombres, l'équation (10.10) fournit une estimation issue d'un estimateur convergent (quand le nombre de particules tend vers l'infini) de la quantité recherchée :

$$P_\lambda(X_i(0) = N \mid \mathbf{y}) = \int_{\mathsf{Z}^{m-1}} P_\lambda(X_i(0) = N \mid Z_{1:m-1} = z_{1:m-1}, \mathbf{y}) dP_\lambda(Z_{1:m-1} = z_{1:m-1} \mid \mathbf{y})$$

Avec cette méthode, on a seulement besoin d'effectuer une seule évolution particulaire pour obtenir une estimation de la loi du nucléotide à la racine en tous les sites i.

10.4.2 Comparaison entre les deux méthodes

On compare dans cette section les deux méthodes d'inférence à la racine proposées dans la section 10.4.1, en particulier la convergence des estimations obtenues (lorsque le nombre de particules tend vers l'infini).

Exemple 10.4.1. *On considère deux modèles, le premier atypique (les taux de sauts sont choisis pour favoriser les phénomènes de dépendance) et le deuxième typique. Ils sont décrits par respectivement (voir l'annexe A pour une description des modèles) :*

$$\lambda_5 = (\mathsf{R}_{iid}, \mathsf{T}_2(0.5), \mathsf{M}_6(1)),$$

$$\lambda_6 = (\mathsf{R}_{iid}, \mathsf{T}_2(0.5), \mathsf{M}_2(1)).$$

Pour pouvoir utiliser directement la première méthode d'inférence à la racine, on remarque que l'on a choisi le modèle à sites indépendants R_{iid} à la racine.

Comparaison de l'écart quadratique moyen. Pour des séquences courtes de longueurs 3 ou 5, on peut comparer les probabilités obtenues par les simulations particulaires avec les probabilités exactes obtenus par calcul matriciel.

On considère ici quatre exemples, donnés respectivement par :

1. le modèle λ_5 et les observations $\mathbf{y} = (CTY, TCY)$,
2. le modèle λ_5 et les observations $\mathbf{y} = (CCTTY, TTCCY)$,
3. le modèle λ_6 et les observations $\mathbf{y} = (RCY, RCY)$,
4. le modèle λ_6 et les observations $\mathbf{y} = (CACCY, RGCCY)$.

Pour chaque exemple, on calcule d'une part les probabilités exactes pour que le nucléotide central à la racine soit égal à A, C, G ou T. Pour le premier exemple, on obtient :

$$\begin{aligned} P(X_2(0) = A \mid \mathbf{y}) &= 0.06, \\ P(X_2(0) = C \mid \mathbf{y}) &= 0.21, \\ P(X_2(0) = G \mid \mathbf{y}) &= 0.18, \\ P(X_2(0) = T \mid \mathbf{y}) &= 0.55. \end{aligned}$$

On calcule d'autre part des estimations particulaires de ces valeurs avec la première et la deuxième méthode, pour 100 répétitions de n particules, où $n \in \{100, 1000, 10000, 100000\}$. On en déduit l'écart quadratique moyen pour chaque méthode et chaque n. Pour les exemples 1. et 4., en tenant compte du fait qu'il faut quatre simulations avec la première méthode contre une seule pour la deuxième méthode (en considérant qu'il faut 4 fois plus de particules nécessaires pour obtenir une valeur avec la première méthode), on représente sur les figures 10.18 et 10.19 les écarts quadratiques moyens obtenus en fonction du nombre de particules, affichés sur un repère log-log.

On observe dans tous les cas une décroissance en $1/n$ de l'écart quadratique moyen quand le nombre de particules n augmente. De plus, pour les exemples 1. et 2., la deuxième méthode donne une meilleure précision pour les nucléotides qui sont les plus probables

au site que l'on souhaite inférer (par exemple les probabilités d'obtenir C, G et T sont mieux estimées par la deuxième méthode dans l'exemple 1.). Pour les exemples 3. et 4., la première méthode est beaucoup plus précise que la deuxième méthode dans tous les cas (pour l'exemple 4., il faudrait entre 60 et 1500 fois plus de particules avec la deuxième méthode pour obtenir le même écart, suivant les nucléotides considérés).

Comparaison de l'écart-type. On considère ici deux exemples, donnés par :

1. le modèle λ_5 et les feuilles de longueur 13 :

$$\mathbf{y} = (CCCCCCTTTTTTY, TTTTTTCCCCCCY),$$

2. le modèle λ_6 et les feuilles de longueur 13 :

$$\mathbf{y} = (RGGACACCTGACY, TGAGGGCCCAATA).$$

Pour chaque exemple, on cherche encore à inférer le nucléotide central à la racine (pour l'exemple 1., les observations pour ce site est le couple (T, C)). On ne peut pas calculer exactement les probabilités pour que ce nucléotide à la racine soit égal à A, C, G ou T. On va alors comparer les écarts-types des probabilités obtenues à l'aide des méthodes particulaires. Comme dans le paragraphe précédent, on utilise les deux méthodes et 100 répétitions de $n \in \{100, 1000, 10000, 100000\}$ particules. Dans l'exemple 1., pour les deux méthodes et avec 100000 particules, on obtient les estimations suivantes :

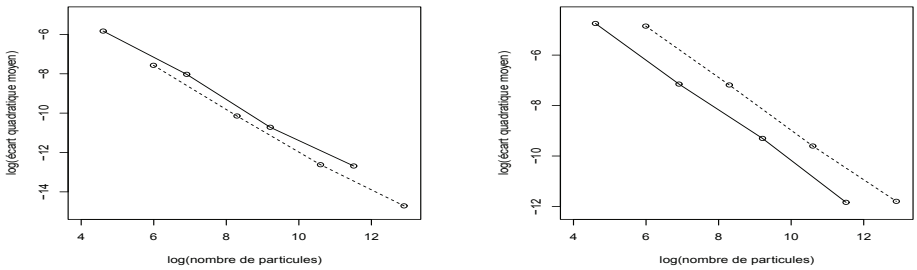
$$0.03 \text{ pour } A,\ 0.14 \text{ pour } C,\ 0.06 \text{ pour } G \text{ et } 0.77 \text{ pour } T.$$

En tenant compte du fait qu'il faut quatre simulations avec la première méthode contre une seule pour la deuxième méthode (ce qui se traduit par considérer 4 fois plus de particules nécessaires pour obtenir une valeur avec la première méthode), on représente sur les figures 10.20 et 10.21 les écarts-types empiriques obtenus en fonction du nombre de particules, sur un repère log-log.

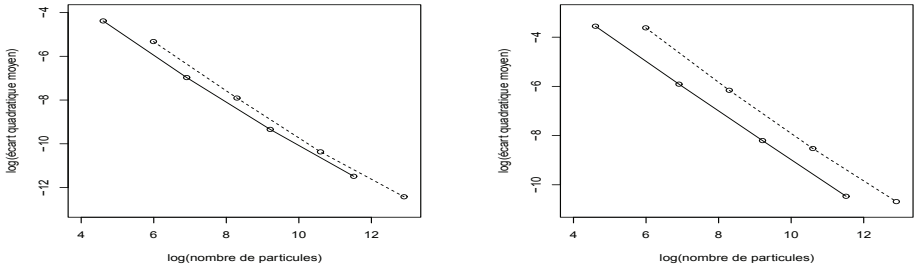
On observe une décroissance en $1/\sqrt{n}$ de l'écart-type. Pour l'exemple 1., les deux méthodes fournissent pour cette longueur de séquence des résultats similaires, avec un écart-type légèrement plus faible pour les nucléotides les plus probables (C et T) avec la deuxième méthode. Pour l'exemple 2., la première méthode est beaucoup plus précise que la deuxième méthode dans tous les cas.

Normalité. Pour tous les exemples des paragraphes précédents, on utilise le test de normalité de Shapiro-Wilk [104], et on obtient dans tous les cas que l'on ne peut pas rejeter l'hypothèse de normalité à risque 1% lorsque l'on utilise 100000 particules (avec 10000 particules, on obtient la même conclusion hormis pour deux échantillons).

Conclusion. On déduit des exemples précédents que les deux méthodes peuvent être utilisées pour effectuer l'inférence à la racine d'un site, et que dans certains cas la première méthode fournit un résultat plus précis que la deuxième à coût constant. Néanmoins, la deuxième méthode peut être mise en œuvre directement dans des cadres plus généraux que la première méthode, c'est-à-dire dans le cas où la loi à la racine est la loi stationnaire du modèle d'évolution considéré ou une approximation markovienne de celle-ci. De plus, elle permet en une seule simulation de réaliser l'inférence de tous les nucléotides de la séquence. C'est pourquoi dans la suite de ce chapitre on utilise uniquement la deuxième méthode d'inférence de nucléotides à la racine.

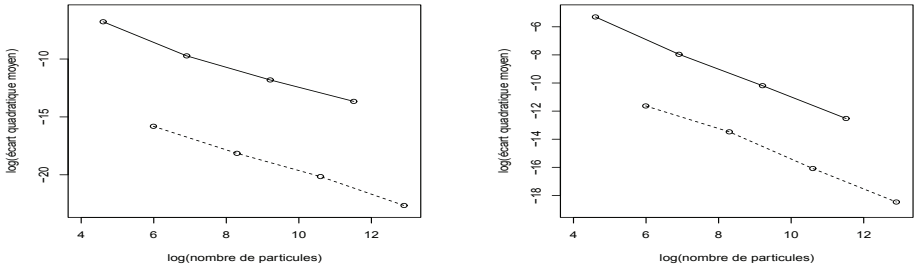


(a) A (b) C

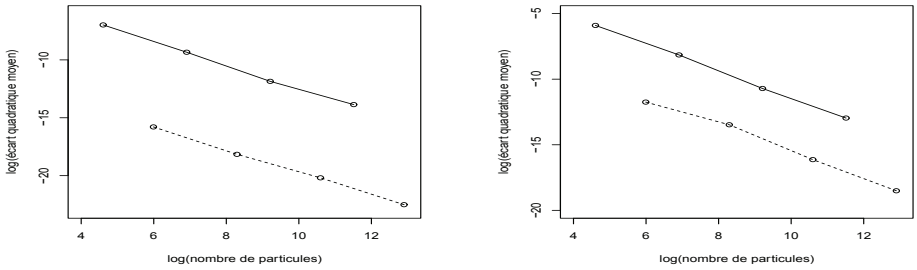


(c) G (d) T

FIGURE 10.18 – Modèle atypique à 3 nucléotides.



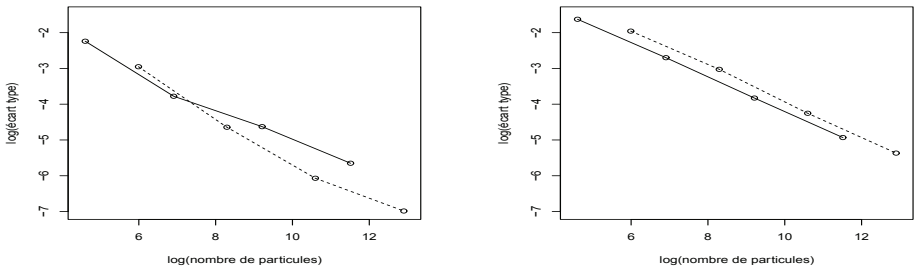
(a) A (b) C



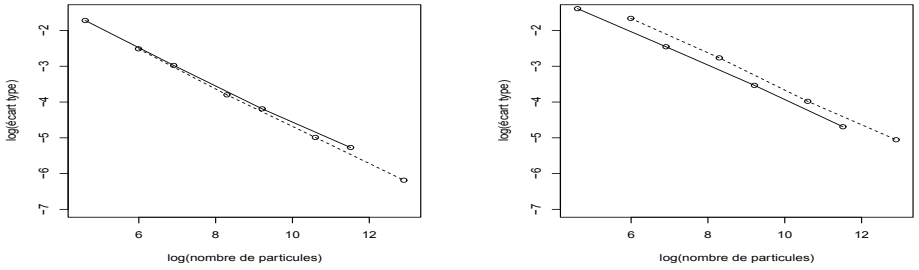
(c) G (d) T

FIGURE 10.19 – Modèle typique à 5 nucléotides.

Évolution du logarithme de l'écart quadratique moyen de la probabilité que le nucléotide central à la racine soit égal à A, C, G ou T, en fonction du logarithme du nombre total de particules utilisées, pour la méthode une (en pointillés) et pour la méthode deux (ligne pleine noire). Les modèles sont décrits dans la section 10.4.2.

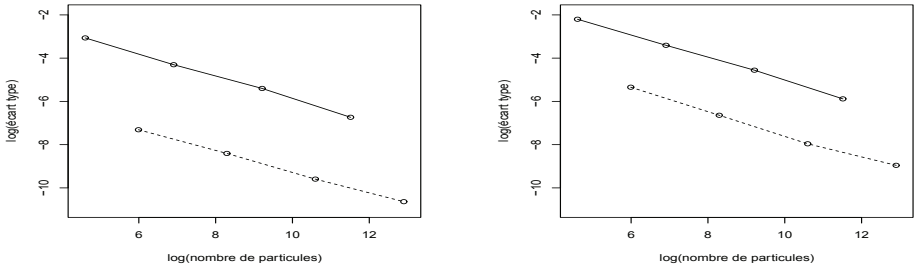


(a) A (b) C

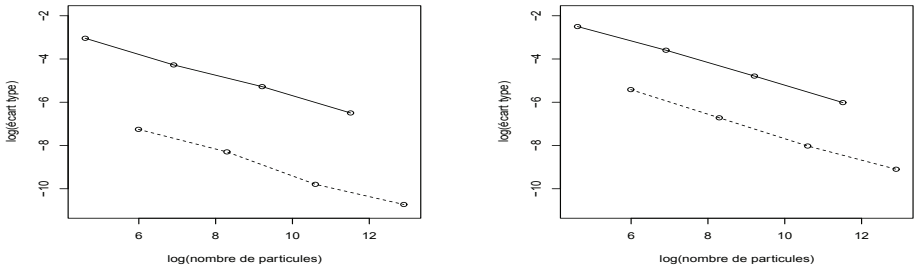


(c) G (d) T

FIGURE 10.20 – Modèle atypique à 13 nucléotides.



(a) A (b) C



(c) G (d) T

FIGURE 10.21 – Modèle typique à 13 nucléotides.

Évolution du logarithme de l'écart-type de la probabilité que le nucléotide central à la racine soit égal à A, C, G ou T, en fonction du logarithme du nombre total de particules utilisées, pour la méthode une (en pointillés) et pour la méthode deux (ligne pleine noire). Les modèles sont décrits dans la section 10.4.2.

10.4.3 Influence de la séquence observée - modèle atypique

On veut étudier dans cette section la distance jusqu'à laquelle il est pertinent de prendre en compte les sites voisins pour estimer le nucléotide ancestral en un site, sur un modèle atypique où la dépendance aux voisins est forte.

On va pour cela considérer des observations constituées de séquences de longueur m, puis effectuer l'inférence du nucléotide central à la racine avec des observations restreintes.

Précisément, pour une séquence $\mathbf{s}$ de longueur impaire m, le nucléotide central est situé en position $\lceil m/2 \rceil$ (la séquence est numérotée selon $[\![1, m]\!]$). On définit alors pour tout nombre impair $0 \leq k \leq m$ la séquence restreinte de longueur k comme $\mathbf{s}[\![\lceil m/2 \rceil - \lfloor k/2 \rfloor, \lceil m/2 \rceil + \lfloor k/2 \rfloor]\!]$.

Modèle et observations considérés. On choisit un modèle où les dépendances aux voisins sont fortes et on reprend donc le modèle atypique

$$\lambda_5 = (\mathsf{R}_{iid}, \mathsf{T}_2(0.5), \mathsf{M}_6(1)).$$

Les trois exemples considérés correspondent à un changement des observations, choisies pour favoriser les dépendances.

Exemple 10.4.2. *On utilise les observations :*

$$\mathbf{y} = (CCCCCCCCTTTTTTTTT, TTTTTTTTCCCCCCCCC).$$

Pour les restrictions de ces observations de longueurs 3 *à* 17*, on cherche à inférer le nucléotide central à la racine (les observations pour ce site est le couple* (T, C)*). On effectue l'inférence de la même manière que dans la section 10.4.2 à l'aide des méthodes particulaires, avec la deuxième méthode. On observe sur la figure 10.22 les diagrammes en boîte des probabilités d'obtenir* A*,* C*,* G *ou* T*.*

On observe que la probabilité d'obtenir T *augmente de* 0.2 *lorsque l'on considère deux pas de dépendance au lieu d'un seul. Au pas suivant, sa probabilité diminue de* 0.08*. On observe aussi une suite oscillante, comme déjà observée dans l'exemple 2 extrême de la section 5.1.4.*

Exemple 10.4.3. *On utilise les observations :*

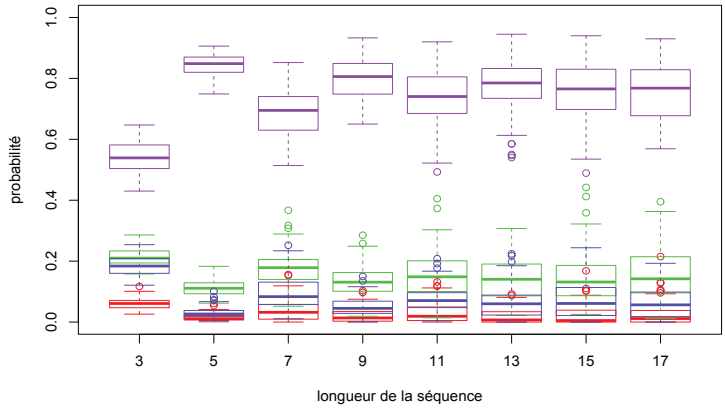
$$\mathbf{y} = (GGAAGGAAGGAAGGAAG, AGAGAGAGAGAGAGAGA).$$

Les observations pour le site central est le couple (G, A)*. On procède comme pour l'exemple précédent et on observe sur la figure 10.23(a) les diagrammes en boîte des probabilités d'obtenir* A*,* C*,* G *ou* T *pour* 100000 *particules. On observe pour cet exemple qu'il est nécessaire d'effectuer l'inférence à la racine avec au moins deux pas de dépendance pour connaître le nucléotide à la racine le plus probable* G*.*

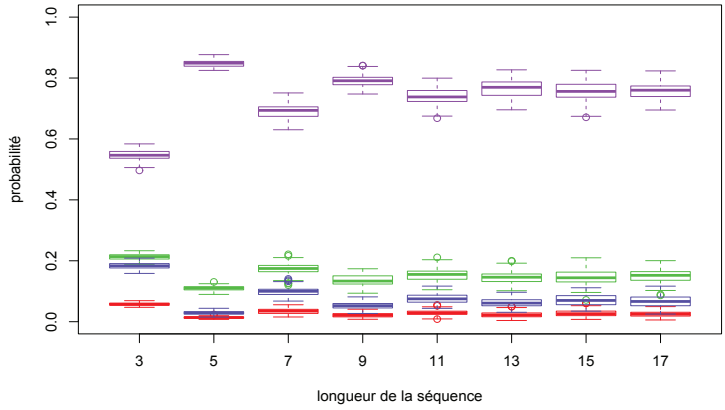
Exemple 10.4.4. *On utilise les observations :*

$$\mathbf{y} = (GGGGGGGAGGAAAAAAA, AAAAAAAGAGAGGGGGG).$$

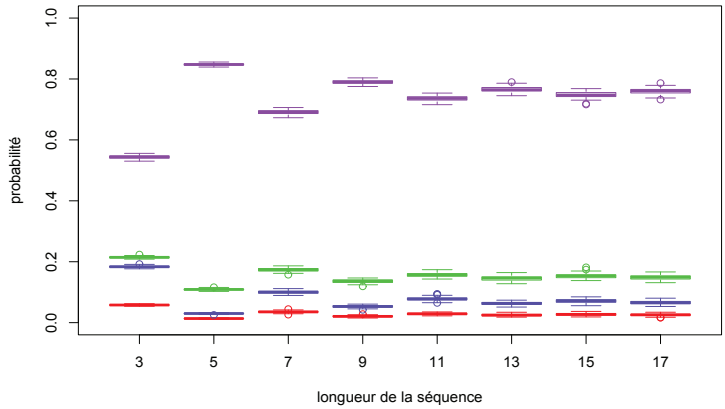
Les observations pour le site central est de nouveau (G, A)*. On procède comme pour le précédent exemple et on observe les diagrammes en boîte sur la figure 10.23(b), pour* 100000 *particules. On observe pour cet exemple une compétition entre* A *et* G *pendant un nombre de pas de dépendance élevé.*



(a) 1000 particules

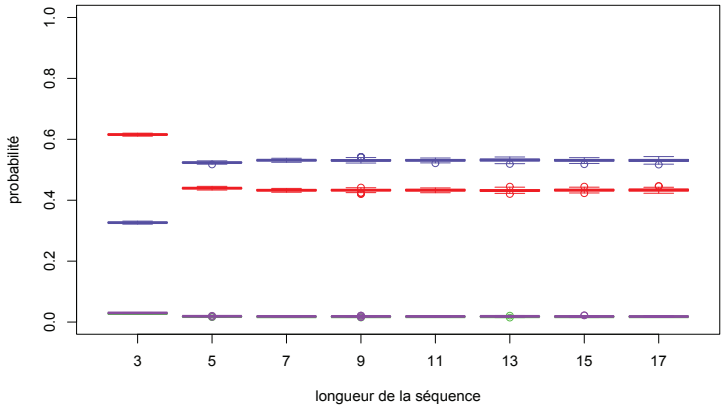


(b) 10000 particules

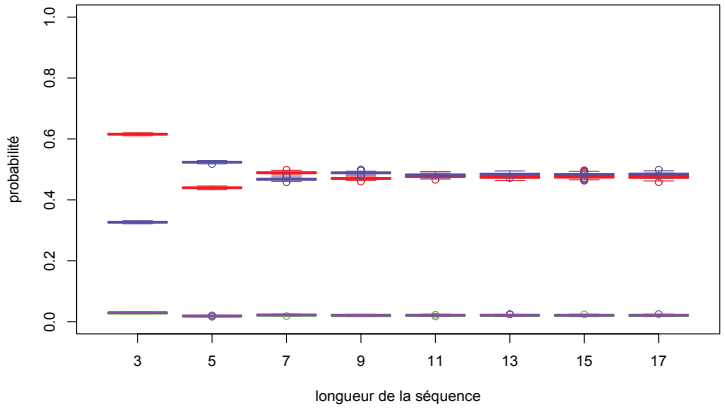


(c) 100000 particules

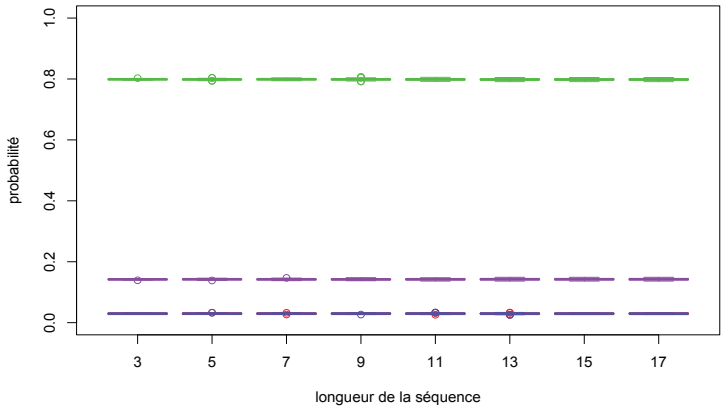
FIGURE 10.22 – Diagrammes en boîte des probabilités d'obtenir A (en rouge), C (en vert), G (en bleu) ou T (en violet) au milieu de la séquence associée à la racine, pour des restrictions symétriques de longueur 3 à 17 et pour l'exemple 10.4.2.



(a) exemple 10.4.3 atypique.



(b) exemple 10.4.4 atypique.



(c) exemple 10.4.5 typique.

FIGURE 10.23 – Diagrammes en boîte des probabilités d'obtenir A (en rouge), C (en vert), G (en bleu) ou T (en violet) au milieu de la séquence associée à la racine, pour des restrictions symétriques de longueur 3 à 17.

Dans ce dernier exemple, le nombre de pas de dépendance à considérer pour obtenir le nucléotide central à la racine le plus probable est 6 (à gauche et à droite). Les méthodes exactes sont ici trop coûteuses pour être utilisées et les méthodes par troncature à 1 ou 2 pas éliminent une part trop importante de la dépendance présente. Les méthodes particulaires décrites sur la structure markovienne de l'évolution permettent ici d'obtenir des estimations précises et en temps raisonnable des différentes probabilités.

10.4.4 Influence de la séquence observée - modèles typiques

On considère dans cette section un modèle d'évolution typique. On observe tout d'abord pour un couple d'observations particulier l'influence de la séquence observée sur l'inférence du nucléotide central à la racine, avant de considérer tous les triplets et quintuplets encodés comme observations.

Le modèle global d'évolution choisi est le modèle typique suivant déjà utilisé dans l'exemple 10.4.1 :

$$\lambda_6 = (\mathsf{R}_{iid}, \mathsf{T}_2(0.5), \mathsf{M}_2(1)).$$

Exemple 10.4.5. *On utilise les observations :*

$$\mathbf{y} = (GGGACACCTGACC, TGAGGGCCCAATA).$$

Comme pour les trois exemples atypiques, on cherche à inférer le nucléotide central à la racine (les observations pour ce site est le couple (C, C)). On observe sur la figure 10.23(c) les diagrammes en boîte des probabilités d'obtenir A, C, G ou T, avec la deuxième méthode utilisée avec 100000 *particules et* 100 *répétitions. On observe qu'un seul pas de dépendance suffit pour inférer correctement les différentes probabilités. En effet, l'écart de probabilité maximal entre considérer un pas ou deux pas de dépendance pour l'inférence d'un nucléotide est* 0.00006.

Plus généralement, on cherche à savoir si pour certaines observations, les dépendances à plus d'un pas ne peuvent pas être négligées, même pour ce modèle typique. On considère alors l'ensemble des couples de quintuplets et on calcule pour chaque couple les différentes probabilités d'inférence du nucléotide au milieu de la racine. Pour chacun de ces couples, on calcule également les probabilités d'inférence pour la restriction aux triplets.

On peut alors calculer pour chaque couple de quintuplet l'écart de probabilité maximal (sur l'ensemble des quatre nucléotides) entre considérer un pas ou deux pas de dépendance. On obtient que l'écart le plus important est obtenu pour le couple $(CATAG, TCTCG)$, avec une différence de probabilité d'un peu plus de 3% d'obtenir le nucléotide T (sa probabilité est de 0.79 pour le couple de triplets contre 0.76 pour le couple de quintuplets).

Le diagramme en boîte pour l'ensemble des couples est donné sur la figure 10.24. On obtient que la médiane est 0.0016 et plus de 96% des couples ont une différence de probabilité plus petite qu'un pour cent.

Pour ce modèle typique, l'influence des voisins non immédiats pour l'inférence d'un nucléotide à un site donné est donc limitée et les méthodes par troncature à 1 ou 2 pas fournissent ici une bonne approximation.

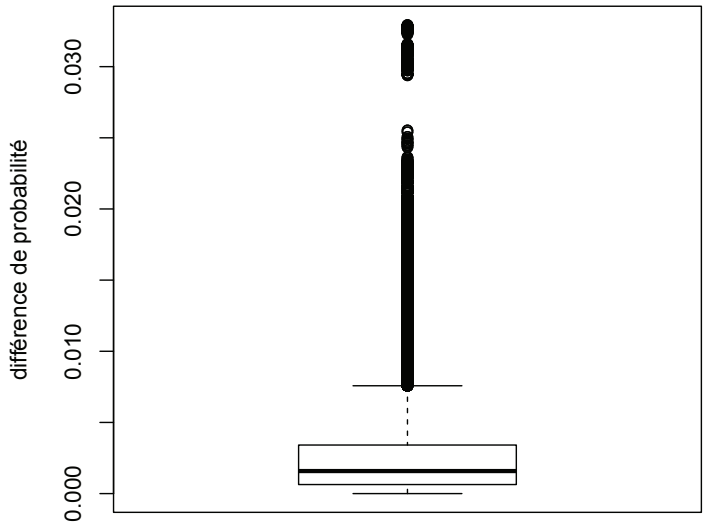


FIGURE 10.24 – Diagramme en boîte de l'écart de probabilité maximal entre considérer un pas ou deux pas de dépendance, pour l'ensemble des couples de quintuplets, pour le modèle de la section 10.4.4.

10.4.5 Inférence de la séquence complète à la racine

On veut dans cette section inférer une séquence complète à la racine (qui peut être de longueur importante), avec la loi à la racine étant fixée comme la loi stationnaire du modèle d'évolution, une approximation markovienne de cette loi stationnaire ou une loi définie indépendamment sur chaque site. Comme l'utilisation de méthodes particulaires est coûteux en temps de calcul, on essaye de privilégier la méthode par troncature à un ou deux pas. On sait néanmoins d'après la section précédente 10.4.3 que si une dépendance à portée longue existe, alors cette méthode par troncature peut donner des résultats erronés.

Pour un site i, on cherche donc d'abord à vérifier si une dépendance à portée longue existe en calculant les probabilités pour que le nucléotide à la racine soit égal à A, C, G ou T, en considérant d'une part la séquence restreinte de longueur 3 et d'autre part la séquence restreinte de longueur 5. Si le nucléotide le plus probable obtenu est le même dans les deux cas, alors on choisit celui-ci comme nucléotide à la racine au site i.

Dans le cas contraire, on utilise une méthode particulaire sur le tronçon de séquence Φ-encodée minimal associé à ce site (voir le corollaire 3.5.2 sur le découpage RY, et la remarque 3.5.4 qui indique que ce découpage reste valide avec la loi à la racine choisie) avec un nombre de particules suffisant pour distinguer significativement le nucléotide le plus probable.

On obtient l'algorithme suivant pour inférer la séquence complète à la racine.

Algorithme 10.4.6. *Pour chaque site i, pour les séquences restreintes de longueur 3 et 5 autour de i, calculer les probabilités pour que le nucléotide à la racine soit égal à A, C, G ou T.*

- *Si le nucléotide le plus probable obtenu est le même dans les deux cas, alors choisir ce nucléotide pour le site i.*
- *Sinon, utiliser la méthode 2 décrite dans la section 10.4.1 utilisant les évolutions particulaire.*

On obtient alors une séquence complète qui est choisie pour inférer la séquence associée à la racine.

10.5 Comparaison des approximations du maximum de vraisemblance

On compare dans cette section différents estimateurs de vraisemblance lorsqu'il s'agit d'estimer le maximum de vraisemblance vis-à-vis d'un paramètre réel. On utilise :

- les estimateurs par couples et triplets encodés, qui fournissent une estimation consistante du paramètre recherché lorsque le nombre de site considéré m tend vers l'infini (voir proposition 4.1.3),
- l'estimateur par approximation markovienne à un pas,
- les estimateurs particulaires, dont les estimations convergent vers celles de l'estimateur du maximum de vraisemblance lorsque le nombre de particules tend vers l'infini (voir théorème 8.2.5).

On montre que les estimateurs particulaires sont difficiles à utiliser efficacement car coûteux en temps de calcul, et que les estimateurs composites fournissent alors une alternative efficace pour estimer le maximum de vraisemblance.

Dans le premier exemple de la section 10.5.1, on choisit un modèle fixé typique et différentes séquences simulées issues de ce modèle (et donc typiques également). On cherche alors à estimer un des paramètres de taux de sauts du modèle connaissant tous les autres paramètres, pour chaque séquence simulée. On utilise pour cela les estimateurs particulaires de la vraisemblance (voir section 8.2) ainsi que les vraisemblances composites obtenues par triplets encodés ou par approximation markovienne (voir chapitre 4). On cherche ensuite à comparer les différentes estimations du maximum de vraisemblance obtenues.

Au niveau de précision considéré (pas de 0.1), on observe que les écarts entre les différents estimateurs du maximum de vraisemblance sont négligeables et qu'en particulier, l'estimateur par approximation markovienne à un pas peut être utilisé.

Dans le deuxième exemple de la section 10.5.1, on montre que les maximums de vraisemblances obtenus avec les estimateurs particulaires et l'estimateur par approximation markovienne à un pas sont différentes en général. On choisit pour cela une séquence atypique et un modèle atypique dont on montre que les maximums de vraisemblance sont estimés de façon significativement différente.

Dans la section 10.5.2, on compare avec une plus grande précision les estimations composites du maximum de vraisemblance obtenues par approximation markovienne à un pas,

par couples encodés et par triplets encodés sur un modèle typique. On en conclut que les trois estimateurs peuvent être utilisés et ont des comportements similaires.

En général, il est difficile de calculer avec une bonne précision une estimation du maximum de vraisemblance avec les estimateurs particulaires puisque le temps de calcul devient long quand le nombre de particules augmente. De plus, l'amplitude de la log-vraisemblance est faible lorsque les coefficients considérés sont proches du maximum de vraisemblance. Les estimateurs particulaires peuvent donc être utilisés pour trouver des régions où la log-vraisemblance est plus importante mais pas pour estimer précisément le maximum de vraisemblance. Ainsi, d'après les résultats de consistance de la proposition 4.1.3, on préfère utiliser l'estimateur par couples ou par triplets encodés pour estimer le maximum de vraisemblance.

Enfin, dans la section 10.5.3, on estime l'écart-type de l'estimateur du maximum de vraisemblance associé à la vraisemblance composite par triplets encodés. On utilise deux estimations différentes, que l'on compare : d'une part l'estimation empirique directe et d'autre part l'estimation semi-empirique proposée dans la section 4.1.2.

10.5.1 Comparaison entre les approximations composites et particulaires

On compare sur un modèle typique (associé à des jeux de séquences typiques) et un modèle atypique (associé à un jeu de séquences atypiques) les estimations du maximum de vraisemblance obtenues par les estimateurs particulaires et par les estimateurs composites.

Données.

Exemple 10.5.1. *(modèle typique). On utilise le modèle* $(\mathsf{R}_{iid}, \mathsf{T}_{10}, \mathsf{M}_2(1))$*. On simule [90] selon ce modèle* 100 *séquences de longueur* 1000*.*

Pour ces séquences simulées, on cherche à retrouver la valeur du paramètre $r_{CG\to TG}$ *connaissant tous les autres paramètres. De façon équivalente, cela correspond à trouver le modèle utilisé pour la simulation parmi l'ensemble* $(\mathsf{M}_2(\theta))_{\theta\in\mathbb{R}^+}$*, le vrai paramètre étant* $\theta_0 = 1$*.*

Exemple 10.5.2. *(modèle atypique). On considère l'ensemble des modèles globaux d'évolution* $\lambda^\theta = (\mathsf{R}_{iid}, \mathsf{T}_1, \mathsf{M}_6(\theta))_{\theta\in\mathbb{R}_*^+}$*. Le paramètre du modèle est ici le coefficient* v_G*. On suppose avoir observé pour un modèle* λ^{θ_0} *(pour un certain* θ_0*) les séquences de longueur* 17 *suivantes :*

$$(CCCCCCCCTTTTTTTTT, TTTTTTTTCCCCCCCCC).$$

On cherche alors à estimer la valeur du paramètre θ_0 *par maximum de vraisemblance, qui correspond ici à trouver la valeur du paramètre* v_G*.*

Méthode et résultats pour le modèle typique. Pour chacune des cent observations provenant de l'exemple 10.5.1, on utilise le corollaire 3.5.2 pour découper la séquence en morceaux indépendants et on utilise sur chaque morceau les estimateurs de vraisemblance suivants, pour chaque paramètre $\theta \in [0, 4]$ et par pas de 0.1 :

- Pour $n = 10$ et avec 100 répétitions, $\hat{L}_{n,\text{1-partic}}$ méthode particulaire avec rééchantillonnage,
- $\hat{L}_{\text{1-Markov}}$ approximation markovienne à 1 pas,
- $\hat{L}_{\text{couples}}$ estimateur par couples encodés (voir remarque 4.1.4),
- $\hat{L}_{\text{triplets}}$ estimateur par triplets encodés (voir définition 4.1.1).

On en déduit alors pour chaque observation et pour chaque estimateur considéré une estimation du maximum de vraisemblance associé.

On observe sur la figure 10.25 les estimations de vraisemblances obtenues pour deux séquences et pour les estimateurs $\hat{L}_{10,\text{1-partic}}$ et $\hat{L}_{\text{1-Markov}}$. Pour les cent observations considérées, les estimations de log-vraisemblance entre les deux méthodes sont proches et aboutissent à des estimations du maximum de vraisemblance identiques ou distantes de 0.1 (avec une précision au dixième). Les estimations $\hat{L}_{\text{couples}}$ et $\hat{L}_{\text{triplets}}$ fournissent également les mêmes estimations du maximum de vraisemblance (avec une précision au dixième).

Remarque 10.5.3. *Autour de chaque coefficient estimé, on recommence le calcul des estimations de vraisemblance avec cette fois un pas de* 0.01 *et* 100 *particules. Les méthodes particulaires fournissent dans ce cas des estimations peu précises et ne permettent pas d'estimer le maximum de vraisemblance à ce niveau de précision.*

La racine carré de l'écart quadratique moyen entre la vraie valeur du paramètre et les estimations du maximum de vraisemblance (avec l'un des quatre estimateurs) est 0.13. Celle entre les estimations du maximum de vraisemblance par l'estimateur particulaire et par approximation markovienne à un pas est 0.06. On en déduit que les estimations de maximum de vraisemblance par estimateur particulaire et par approximation markovienne sont similaires.

Méthode et résultats pour le modèle atypique. Pour l'exemple 10.5.2, on utilise les estimateurs de vraisemblance suivants pour chaque paramètre $\theta \in [0, 4]$ et par pas de 0.1 :

- Pour $n = 1000000$ et avec 100 répétitions, $\hat{L}_{n,\text{1-partic}}$ méthode particulaire avec réechantillonnage,
- $\hat{L}_{\text{1-Markov}}$ approximation markovienne à 1 pas,

On en déduit une estimation du maximum de vraisemblance associé à chacun des deux estimateurs.

On représente sur la figure 10.26 les estimations de log-vraisemblance obtenues. On observe que les estimations sont significativement différentes (ce que l'on a déjà vu dans les sections 10.1.2 et 10.2.1) mais également les estimations du maximum de vraisemblance.

10.5.2 Comparaison entre les approximations composites

Dans l'exemple 10.5.1, en considérant uniquement les vraisemblances composites par couples encodés, triplets encodés et approximation markovienne à un pas, on peut augmenter la précision des estimations du maximum de vraisemblance, pour pouvoir comparer plus précisément les écarts quadratiques moyens. On estime la vraisemblance en chaque coefficient par pas de 0.001 pour chaque estimateur composite. Dans le tableau suivant on résume les différentes racines d'écarts quadratiques moyens :

	θ_0	$\hat{L}_{\text{triplets}}$	$\hat{L}_{\text{couples}}$	$\hat{L}_{\text{1-Markov}}$
θ_0	0	0.123	0.124	0.123
$\hat{L}_{\text{triplets}}$	0.123	0	0.002	0.003
$\hat{L}_{\text{couples}}$	0.124	0.002	0	0.005
$\hat{L}_{\text{1-Markov}}$	0.123	0.003	0.005	0

On observe sur le tableau précédent que les trois estimateurs peuvent être utilisés et ont des comportements similaires pour estimer la maximum de vraisemblance.

10.5.3 Exemple d'estimation de la variance associée aux triplets encodés

On s'intéresse ici au maximum de vraisemblance obtenu par triplets encodés. On cherche à estimer variance de l'estimateur du maximum de vraisemblance de deux manières différentes : d'une part en utilisant l'estimation empirique directe et d'autre part avec l'estimation semi-empirique décrite dans la section 4.1.2. On considère pour cela deux modèles, présentés dans les exemples 10.5.4 et 10.5.5.

Exemple 10.5.4. *On considère pour $\theta \in \mathbb{R}^+$ les modèles*

$$\lambda_1(\theta) = (\mathsf{R}_{iid}, \mathsf{T}_1, \mathsf{M}_1(\theta))$$

définis par (voir l'annexe A pour la définition de tous les modèles) :

- R_{iid} *loi à la racine telle que chaque nucléotide est indépendant des autres et selon la loi :*
$$(0.25, 0.25, 0.25, 0.25).$$
- T_1 *l'arbre constitué de deux arêtes de longueurs* 0.5 *et* 0.6.
- $\mathsf{M}_1(\theta)$ *décrit par :*

$$\begin{aligned}
&v_A = 0.05, v_C = 0.125, v_G = 0.15, v_T = 0.175,\\
&w_A = 0.05, w_C = 0.25, w_G = 0.15, w_T = 0.35,\\
&r_{CG\to CA} = 0, r_{CA\to CG} = 0, r_{TA\to TG} = 1, r_{TG\to TA} = 0,\\
&r_{CA\to TA} = 0, r_{CG\to TG} = \theta, r_{TA\to CA} = 0, r_{TG\to CG} = 0.
\end{aligned}$$

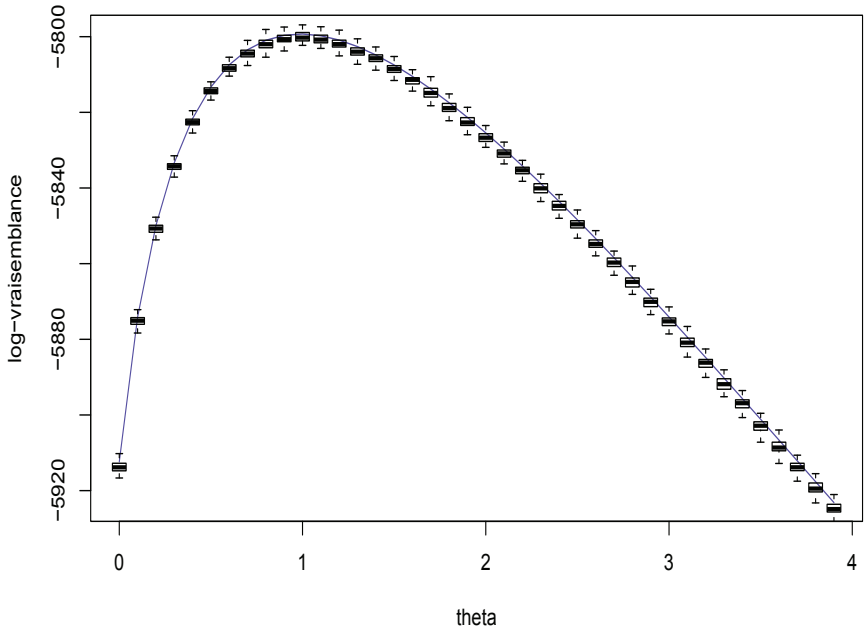
On simule 100 *séquences de longueurs $m = 5000$ selon le modèle $\lambda_1(\theta_0)$ avec $\theta_0 = 2$.*

Pour chaque séquence $k \in [\![1, 100]\!]$ et chaque site $i \in [\![1, m-2]\!]$, on note $\hat{\theta}_0^{i,k}$ l'estimation de θ_0 pour la séquence k le long des sites 1 à i, obtenu avec par maximum de vraisemblance composite triplets par triplets.

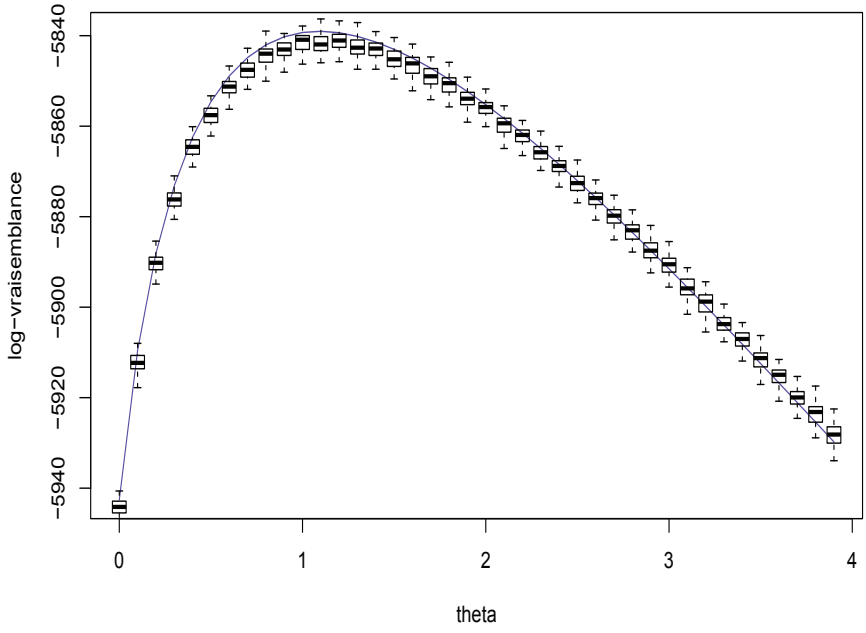
Ensuite, pour tout $i \in [\![1, m-2]\!]$:

- d'une part on calcule l'écart-type empirique de $\left(\hat{\theta}_0^{i,k} - \theta_0\right)_{k\in[\![1,100]\!]}$ multiplié par $\sqrt{i}$.
- d'autre part pour tout $k \in [\![1, 100]\!]$, on utilise le théorème 4.1.7 pour estimer l'écart-type de l'estimateur $\sqrt{i}\hat{\theta}_0^i$.

Sur la figure 10.27, on a représenté le nombre de triplets considérés en abscisses (parmi $[\![1, m-2]\!]$) et l'écart-type estimé en ordonnées. La courbe rouge pleine correspond à l'écart-type estimé empiriquement. La courbe bleue correspond à la moyenne des 100 estimations de l'écart-type obtenues semi-empiriquement. Les courbes en pointillés correspondent aux intervalles de confiance à 95% associés (obtenus par *bootstrap* avec 100 répétitions pour l'estimation empirique).

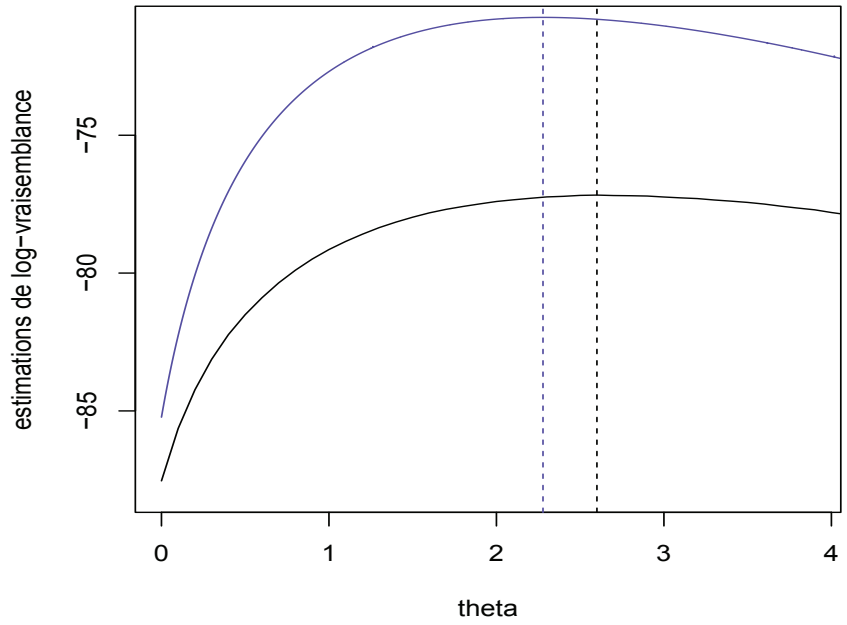


(a) Séquence 1.

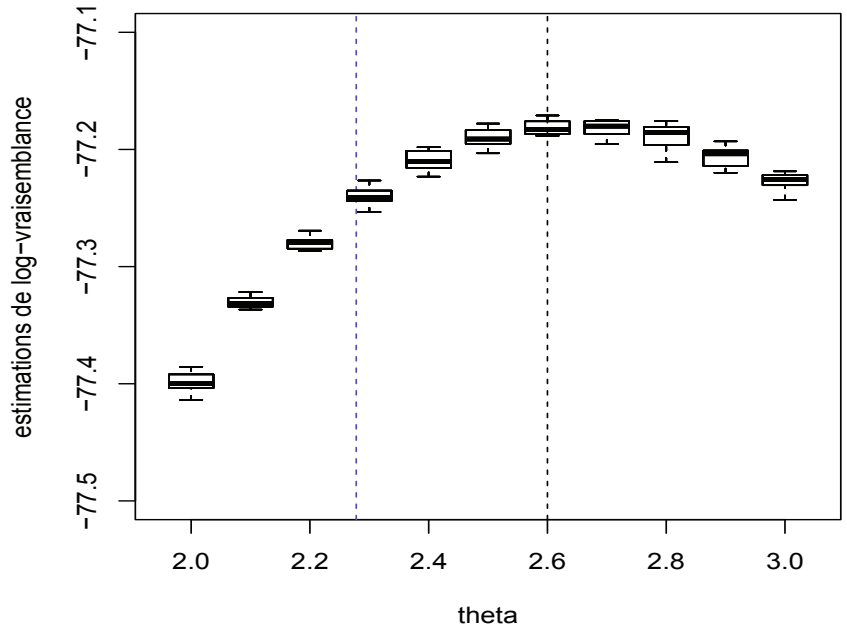


(b) Séquence 4.

FIGURE 10.25 – Estimations de la log-vraisemblance en fonction du coefficient θ pour deux séquences de l'exemple 10.5.1 de longueur 1000. La ligne bleue correspond à l'approximation markovienne à un pas tandis que les diagrammes en boîte correspondent aux estimations particulaires avec rééchantillonnage systématique et 10 particules.



(a) Estimations pour $\theta \in [0, 4]$.



(b) Zoom sur l'intervalle $[2, 3]$.

FIGURE 10.26 – Pour l'exemple 10.5.2, estimations de la log-vraisemblance en fonction du coefficient θ. La ligne bleue correspond à l'approximation markovienne à un pas tandis que la ligne noire et les diagrammes en boîte correspondent aux estimations particulaires avec rééchantillonnage systématique et un million de particules. Les pointillés correspondent aux maximum de vraisemblance obtenus.

Exemple 10.5.5. *On considère pour $\theta \in \mathbb{R}^+$ les modèles :*

$$\lambda_2(\theta) = (\mathsf{R}_{\mathsf{M}_2(1)}, \mathsf{T}_4, \mathsf{M}_2(\theta)).$$

On simule 300 *séquences de longueurs $m = 5000$ selon le modèle $\lambda_2(\theta_0)$ avec $\theta_0 = 1$.*

On procède de la même façon que dans l'exemple précédent, le résultat étant représenté sur la figure 10.28.

Pour les deux exemples, on obtient qu'avec la même quantité d'information on estime de façon plus précise la variance de l'estimateur du maximum de vraisemblance composite triplets par triplets. Notons que la méthode semi-empirique est beaucoup plus coûteuse en temps de calcul, et que l'on pourrait comparer les estimations à temps constant.

10.6 Comparaison de modèles

On souhaite comparer pour des jeux de séquences données la vraisemblance au maximum de vraisemblance pour différentes classes de modèles d'évolution. On cherche ici à maximiser la vraisemblance à la fois sur les paramètres d'évolution des classes de modèles considérées mais également sur la topologie de l'arbre et sur les différentes longueurs de branches.

Dans le but de mettre en évidence le rôle de la prise en compte de l'effet CpG, les classes de modèles d'évolution considérées sont les suivantes : le modèle T92 qui s'inclut dans le modèle RN95, le modèle T92+CpGs qui s'inclut dans le modèle RN95+YpR et le modèle GTR qui ne s'inclut pas dans le modèle RN95 (voir sections 1.1 et 1.2 pour la description de ces modèles, et plus particulièrement la section 1.2.4).

Le modèle T92+CpGs est choisi car c'est le modèle avec le moins de paramètres qui prend en compte la dépendance CpG et distingue les taux transversions et de transitions (voir tableau 1 de [15] pour une comparaison sur deux jeux de séquences de différents modèles inclus dans le modèle RN95+YpR). Le modèle GTR est choisi car c'est le modèle à sites indépendants le plus général qui reste réversible.

On considère les espèces humaine, chimpanzé et macaque, et deux alignements chacun de l'ordre de 2000 nucléotides situés sur le gène HPR. Pour chacun des deux alignements, on estime dans les sections 10.6.1 et 10.6.2 la vraisemblance au maximum de vraisemblance sous la classe de modèles T92+CpGs de différentes manières et on compare les estimations obtenues.

Dans la section 10.6.3, on compare les estimations de vraisemblance au maximum de vraisemblance obtenues pour la classe de modèles T92+CpGs avec celles de T92 et de GTR.

Méthode. Pour les classes de modèles T92 et GTR, qui sont à sites indépendants, on utilise le logiciel `bppml` de Bio++ pour obtenir à partir d'un jeu de séquences les paramètres d'évolution, la topologie et les longueurs de branches de l'arbre qui maximisent la vraisemblance. On obtient directement la vraisemblance associée au maximum de vraisemblance.

Pour la classe de modèles T92+CpGs, le travail précédent et en particulier l'utilisation de l'algorithme 8.2.9 permet d'approcher la vraisemblance des observations dans cette classe de modèles. On rappelle que cet algorithme procède de la façon suivante.

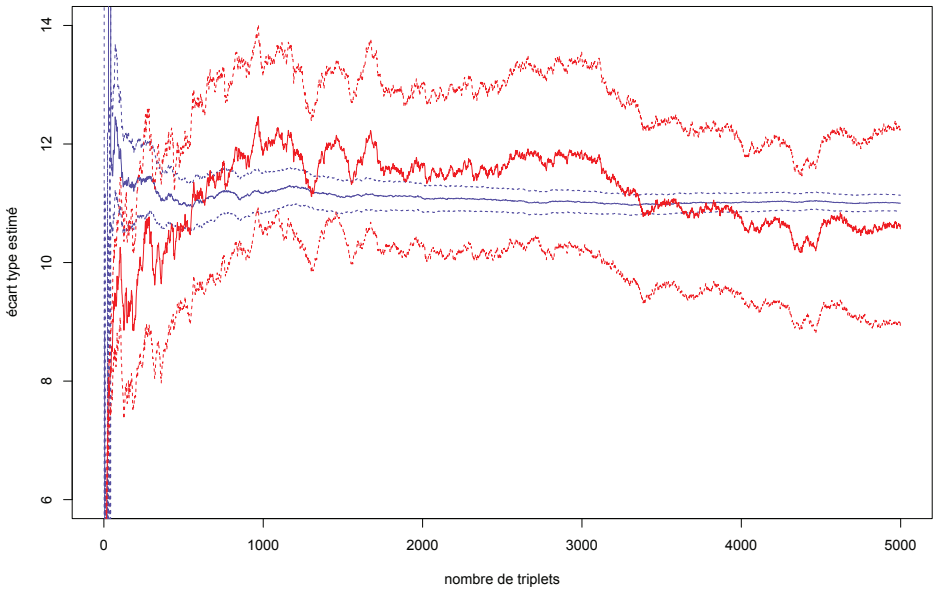


FIGURE 10.27 – Écarts-types de l'estimateur du maximum de vraisemblance par triplets encodés estimés empiriquement (en rouge) ou semi-empiriquement (en bleu), pour l'exemple 10.5.4.

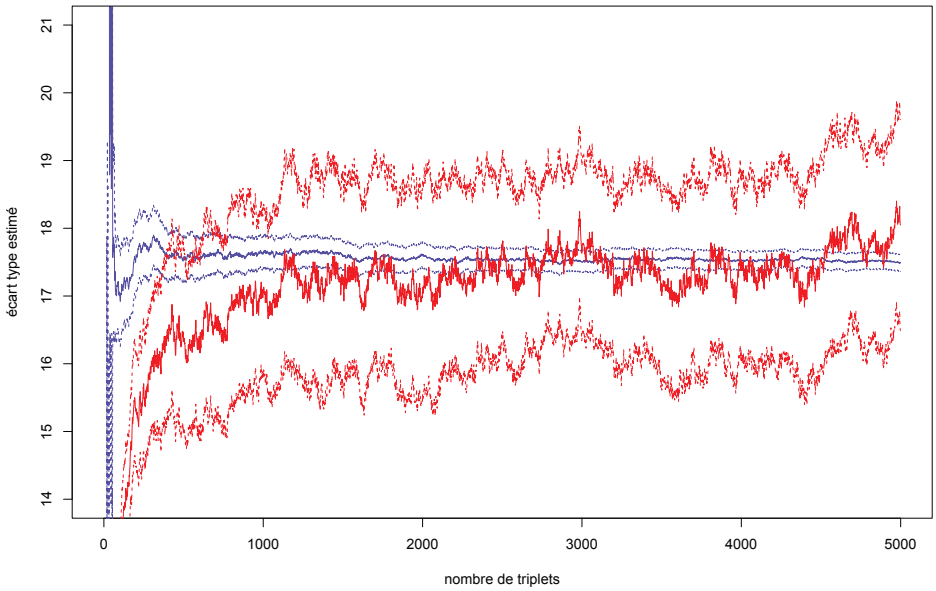


FIGURE 10.28 – Écarts-types de l'estimateur du maximum de vraisemblance par triplets encodés estimés empiriquement (en rouge) ou semi-empiriquement (en bleu), pour l'exemple 10.5.5.

Algorithme 10.6.1. *On choisit un jeu de séquences observés, un estimateur $\hat{L}$ de la vraisemblance et un entier $k \in \{1, 2, 3\}$.*

1. *Estimer $\hat{\theta}_0$ le maximum de vraisemblance (qui est la donnée des paramètres d'évolution, de la topologie et des longueurs de branches de l'arbre) des observations sous le modèle T92+CpGs par la méthode des triplets encodés (voir section 4.1.1).*
2. *Découper les observations en morceaux indépendants grâce au corollaire 3.5.2.*
3. *Pour chaque morceau :*
   - *Si le nombre de nucléotides est inférieur ou égal à $k + 2$, calculer exactement la vraisemblance de ce morceau.*
   - *Sinon, calculer par l'estimateur $\hat{L}$ une estimation de la vraisemblance du morceau.*
4. *Faire le produit des vraisemblances de chaque morceau pour obtenir une estimation de la vraisemblance au maximum de vraisemblance.*

L'utilisation du logiciel `bppml` de Bio++ [15, 38, 39] permet d'effectuer le point 1. de l'algorithme 10.6.1, c'est-à-dire de calculer l'estimateur du maximum de vraisemblance par la méthode des triplets encodés pour le modèle T92+CpGs. Cela fournit la topologie de l'arbre, les différentes longueurs de branches ainsi que les différents paramètres de sauts du modèle.

Pour le point 3. de l'algorithme 10.6.1, on choisit $k \in \{1, 2, 3\}$ et les approximations $\hat{L}$ suivantes (le modèle à la racine est ici choisi comme la chaîne de Markov d'ordre un correspondant à la loi stationnaire pour les triplets encodés) :

- Pour $n \in \{100, 1000, 10000, 100000\}$ et avec 100 répétitions :
  - $\hat{L}_{n,\text{1-partic}}$ méthode particulaire avec rééchantillonnage à chaque pas.
  - $\hat{L}_{n,\text{0-partic}}$ méthode particulaire sans rééchantillonnage.
- $\hat{L}_{k\text{-Markov}}$ approximation markovienne à k pas.

10.6.1 Vraisemblance sous le modèle T92+CpGs pour l'alignement 1

L'alignement est constitué de 2215 nucléotides. On obtient par la première étape de l'algorithme 10.6.1 une estimation du maximum de vraisemblance, donnée par :

- l'arbre et les différentes longueurs de branches suivants (avec la notation de Newick [5]) :

  $$((\text{Homme} : 0.00270724, \text{Chimpanzé} : 0.00350562) : 0.0131641, \text{Macaque} : 0.0115225);$$

- les paramètres de sauts du modèle suivants (voir section 1.1 pour la signification des paramètres κ et θ dans le modèle T92) :

  $$\kappa = 3.61608, \qquad \theta = 0.454953, \qquad r_{CG\to TG} = r_{CG\to CA} = 5.83473.$$

On découpe ensuite en morceaux indépendants grâce au corollaire 3.5.2 et on obtient 499 morceaux de longueurs 2 à 20. Pour chaque nucléotide on associe la longueur du morceau auquel il appartient, et on obtient l'histogramme représenté en figure 10.29.

La proportion de nucléotides qui est calculée de manière exacte suivant la valeur de $k \in \{1, 2, 3\}$ est donnée par le tableau suivant :

k	3	4	5
	25%	42%	56%

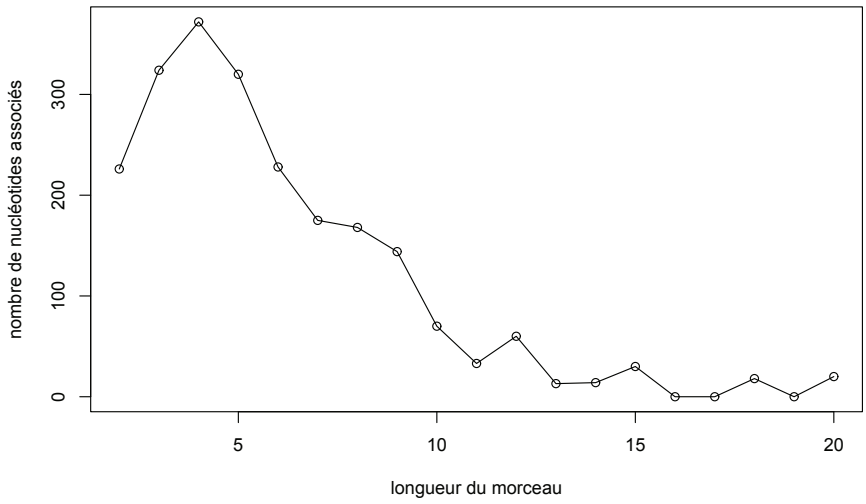


FIGURE 10.29 – Nombre de nucléotides appartenant à un morceau de longueur donnée, pour l'alignement 1.

On calcule ensuite pour les différentes approximations $\hat{L}$ une estimation de la vraisemblance au maximum de vraisemblance pour le modèle T92+CpGs.

On représente dans la figure 10.30 les estimations de log-vraisemblance ainsi que l'intervalle de confiance obtenu pour les différentes approximations $\hat{L}$, respectivement :

- $\hat{L}_{100,1\text{-partic}}$ avec $k = 1, 2, 3$,
- $\hat{L}_{100,0\text{-partic}}$ avec $k = 1, 2, 3$,
- $\hat{L}_{1000,1\text{-partic}}$ avec $k = 1, 2, 3$,
- $\hat{L}_{1000,0\text{-partic}}$ avec $k = 1, 2, 3$,
- $\hat{L}_{10000,1\text{-partic}}$ avec $k = 1, 2, 3$,
- $\hat{L}_{10000,0\text{-partic}}$ avec $k = 1, 2, 3$,
- $\hat{L}_{100000,1\text{-partic}}$ avec $k = 1, 2, 3$,
- $\hat{L}_{100000,0\text{-partic}}$ avec $k = 1, 2, 3$,
- $\hat{L}_{k\text{-Markov}}$ avec $k = 1, 2, 3$.

Les diagrammes en boîte en rouge correspondent à l'utilisation d'un algorithme particulaire avec rééchantillonnage à chaque étape, celles en noir à l'utilisation d'un algorithme particulaire sans rééchantillonnage et les points bleus à l'utilisation de l'approximation markovienne. Numériquement, pour un nombre de particule fixés, l'utilisation de la méthode avec rééchantillonnage systématique requiert environ 50% de temps de calcul supplémentaire.

On obtient des valeurs très proches pour tous les algorithmes utilisés, avec en particulier un bon comportement des méthodes d'approximation markovienne sous ce modèle. Cela fournit une estimation crédible de la vraisemblance sous le modèle T92+CpGs.

10.6.2 Vraisemblance sous le modèle T92+CpGs pour l'alignement 2

On procède comme dans la section précédente, sur un alignement constitué de 2176 nucléotides. On obtient :

– l'arbre et les différentes longueurs de branches suivants :

$$((\text{Homme} : 0.00585431, \text{Chimpanzé} : 0.00783041) : 0.0112768, \text{Macaque} : 0.0406516);$$

– les paramètres de sauts du modèle suivants :

$$\kappa = 4.1433 \quad \theta = 0.522936 \quad r_{CG\to TG} = r_{CG\to CA} = 21.4319.$$

Après découpage en morceaux indépendants, on obtient la même forme de découpage et on calcule ensuite pour les différentes approximations $\hat{L}$ une estimation de la vraisemblance au maximum de vraisemblance pour le modèle T92+CpGs, et on représente sur la figure 10.31 les estimations de log-vraisemblance ainsi que l'intervalle de confiance obtenu pour les différentes approximations $\hat{L}$, dans le même ordre que celui donné dans la section précédente.

10.6.3 Comparaison des vraisemblances obtenues sous trois modèles

On calcule la vraisemblance au maximum de vraisemblance pour les deux alignements sous les modèles T92 et GTR à l'aide du logiciel `bppml` de Bio++. Comme les modèles T92+CpGs et GTR contiennent tous les deux le modèle T92, on sait que l'estimation de la log-vraisemblance va être plus grande dans ces deux premiers modèles. On utilise alors les critères d'information AIC [1] et BIC [103] pour pénaliser les modèles en fonction du nombre de paramètres. En notant k le nombre de paramètres, m le nombre de nucléotides de l'échantillon et L l'estimation de la log-vraisemblance, ces critères s'expriment par :

$$\text{AIC} = -2\log L + 2k \text{ et} \quad \text{BIC} = -2\log L + k\log m.$$

On regroupe dans le tableau suivant ces informations :

modèle	T92	T92+CpGs	GTR
nombre de paramètres	2	3	8
alignement 1 : log-vrais.	−3432	−3389	−3428
alignement 1 : AIC	6868	6784	6872
alignement 1 : BIC	6879	6801	6918
alignement 2 : log-vrais.	−3680	−3580	−3673
alignement 2 : AIC	7364	7166	7362
alignement 2 : BIC	7375	7183	7407

On observe que la prise en compte de la dépendance aux voisins de type CpG améliore significativement la vraisemblance.

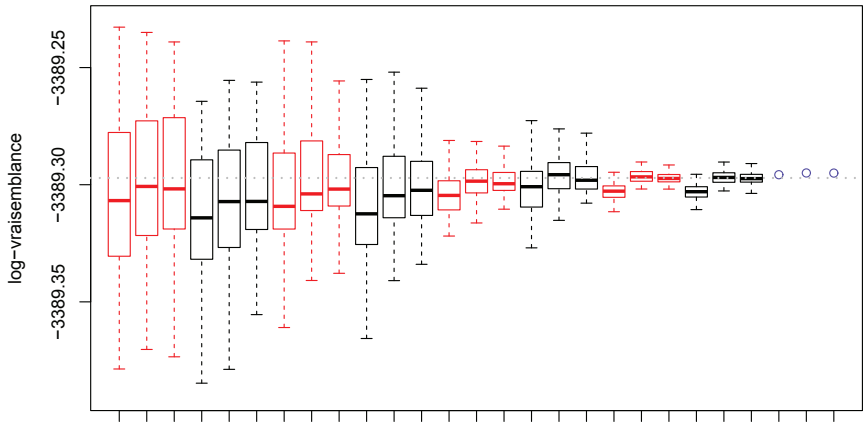


FIGURE 10.30 – Estimations de log-vraisemblance et intervalle de confiance pour l'alignement 1 obtenus pour les différentes approximations $\hat{L}$, dont l'ordre est décrit dans la section 10.6.1. La ligne pointillée grise correspond à la valeur obtenue par $\hat{L}_{100000,\text{1-partic}}$ avec $k = 3$.

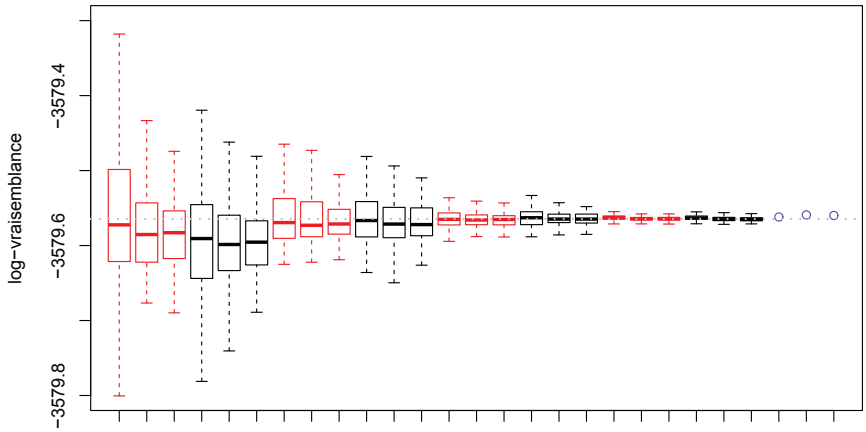


FIGURE 10.31 – Estimations de log-vraisemblance et intervalle de confiance pour l'alignement 2 obtenus pour les différentes approximations $\hat{L}$, dont l'ordre est décrit dans la section 10.6.1. La ligne pointillée grise correspond à la valeur obtenue par $\hat{L}_{100000,\text{1-partic}}$ avec $k = 3$.

Perspectives

Les approximations particulaires développées dans cette thèse permettent d'approcher de façon consistante la vraisemblance exacte d'un modèle issu de la classe avec dépendance RN95+YpR. Elles permettent en outre de quantifier le biais introduit par la vraisemblance composite par approximation markovienne, et donnent donc un moyen de vérifier la précision de cette approximation.

La stratégie d'approximation de la vraisemblance par algorithme particulaire est basée sur la structure de chaîne de Markov cachée décrite dans la section 6.2. Une perspective intéressante serait de développer également des méthodes d'inférence de type EM en exploitant cette structure unilatérale, en adaptant par exemple l'approche de [45].

Il serait aussi intéressant de tirer parti des autres structures de dépendance décrites au chapitre 6, pour construire d'autres méthodes de simulation visant à approcher de façon consistante la vraisemblance d'un modèle RN95+YpR. En particulier, pour la structure de champ markovien d'ordre 1 (présentée dans la section 6.1), une approche MCMC par échantillonnage de Gibbs des histoires évolutives peut être envisagée. On pourrait alors coupler ce type de simulation à un algorithme d'intégration thermodynamique [62] pour approcher pas à pas la vraisemblance du modèle voulu, depuis une condition initiale reposant sur un modèle RN95 sans dépendance (voir [9, 10, 11] pour des exemples d'utilisation d'intégration thermodynamique), et obtenir ainsi des estimations de la vraisemblance qu'il serait intéressant de comparer avec celles produites par l'approche particulaire développée dans cette thèse.

Dans une autre direction, on pourrait envisager d'utiliser des méthodes particulaires pour approcher directement la fonction de gradient et la matrice hessienne de la vraisemblance [30, 97], dans le but d'obtenir un algorithme alternatif de recherche du maximum de vraisemblance, permettant également une estimation de l'erreur associée. Le défi porterait alors sur la réduction du coût de calcul associé à de tels algorithmes.

Enfin, il serait intéressant d'étendre notre approche à des modèles d'évolution plus généraux (voir la discussion dans la section 1.4). En particulier, on pourrait considérer des modèles incorporant des vitesses d'évolution différentes pour chaque site pour accroître le réalisme du modèle (voir par exemple [49, 113, 121]).

Annexe A

Description des modèles

Modèles d'évolution. Les modèles d'évolution suivants sont des modèles d'évolution RN95+YpR. On utilise les notations de la section 1.2 et on les définit par leurs paramètres :

– Pour $\theta \in \mathbb{R}^+$, $\mathsf{M}_1(\theta)$ est décrit par :

$$\begin{aligned}
&v_A = 0.05, v_C = 0.125, v_G = 0.15, v_T = 0.175,\\
&w_A = 0.05, w_C = 0.25, w_G = 0.15, w_T = 0.35,\\
&r_{CG\to CA} = 0, r_{CA\to CG} = 0, r_{TA\to TG} = 1, r_{TG\to TA} = 0,\\
&r_{CA\to TA} = 0, r_{CG\to TG} = \theta, r_{TA\to CA} = 0, r_{TG\to CG} = 0.
\end{aligned}$$

– Pour $\theta \in \mathbb{R}^+$, $\mathsf{M}_2(\theta)$ est décrit par :

$$\begin{aligned}
&v_A = 0.1576, v_C = 0.3234, v_G = 0.3380, v_T = 0.1810,\\
&w_A = 0.4959, w_C = 0.9278, w_G = 1.050, w_T = 0.4000,\\
&r_{CG\to CA} = 1.8942, r_{CA\to CG} = 0, r_{TA\to TG} = 0.3339, r_{TG\to TA} = 0.1263,\\
&r_{CA\to TA} = 0.2570, r_{CG\to TG} = 3.5230\theta, r_{TA\to CA} = 1.9719, r_{TG\to CG} = 0.
\end{aligned}$$

– M_3 est décrit par (les coefficients non indiqués sont égaux à 0.01) :

$$v_G = 10,\ r_{CG\to CA} = 100.$$

– M_4 est décrit par

$$\begin{aligned}
&v_A = 7.199,\ v_C = 6.235,\ v_G = 0.241,\ v_T = 7.313,\\
&w_A = 8.702,\ w_C = 6.914,\ w_G = 7.538,\ w_T = 0.314,\\
&r_{CG\to CA} = 3.821,\ r_{CA\to CG} = 3.363,\ r_{TA\to TG} = 3.340,\ r_{TG\to TA} = 2.517,\\
&r_{CA\to TA} = 5.614,\ r_{CG\to TG} = 8.155,\ r_{TA\to CA} = 8.020,\ r_{TG\to CG} = 7.705.
\end{aligned}$$

– M_5 est décrit par :

$$\begin{aligned}
&v_A = 0.042, v_C = 0.083, v_G = 0.125, v_T = 0.167,\\
&w_A = 0.209, w_C = 0.250, w_G = 0.292, w_T = 0.334,\\
&r_{CG\to CA} = 0.104, r_{CA\to CG} = 0.730, r_{TA\to TG} = 0.438, r_{TG\to TA} = 0.730,\\
&r_{CA\to TA} = 1.501, r_{CG\to TG} = 1.835, r_{TA\to CA} = 1.627, r_{TG\to CG} = 1.877.
\end{aligned}$$

– Pour $\theta \in \mathbb{R}^+$, $\mathsf{M}_6(\theta)$ est décrit par :

$$\begin{aligned}
&v_A = 0.01, v_C = 0.2, v_G = 0.2\theta, v_T = 0.01,\\
&w_A = 0.01, w_C = 0.01, w_G = 0.01, w_T = 0.01,\\
&r_{CG\to CA} = 0.2, r_{CA\to CG} = 0.2, r_{TA\to TG} = 0.2, r_{TG\to TA} = 0,\\
&r_{CA\to TA} = 0, r_{CG\to TG} = 0.2, r_{TA\to CA} = 0.2, r_{TG\to CG} = 1.
\end{aligned}$$

– Pour $\varepsilon > 0$, $\mathsf{M}_{extIrr}(\varepsilon)$ est décrit par (les coefficients non indiqués sont nuls) :

$$v_C = v_G = 1, \quad r_{TG\to CG} = r_{CG\to CA} = r_{CA\to TA} = 1/\varepsilon.$$

$$v_A = v_T = w_A = w_C = w_G = w_T = \varepsilon.$$

Lois à la racine. Les lois à la racine utilisées sont les suivantes :

– R_{iid} vérifie que chaque nucléotide est indépendant des autres et selon la loi :

$$(0.25, 0.25, 0.25, 0.25).$$

– R_{M} correspond à la loi stationnaire d'un modèle d'évolution M.

Arbres. Les arbres considérés sont les suivants (avec la notation de Newick [5]) :

– $\mathsf{T}_1 = (n1 : 0.5, n2 : 0.6)$;
– $\mathsf{T}_2(t) = (n1 : t, n2 : t)$; pour $t > 0$.
– $\mathsf{T}_4 = (Pongo : 0.33636, (GGorilla : 0.17147, (Ppaniscus : 0.19268, Hsapiens : 0.11927) : 0.08386) : 0.06124)$; (représenté sur la figure A.1).
– $\mathsf{T}_6 = (Bovine : 0.69395, (Hylobates : 0.36079, (Pongo : 0.33636, (GGorilla : 0.17147, (Ppaniscus : 0.19268, Hsapiens : 0.11927) : 0.08386) : 0.06124) : 0.15057) : 0.54939)$; (représenté sur la figure A.2)
– $\mathsf{T}_{10} = (((((((hg19 : 0.00485998, panTro2 : 0.00554111) : 0.000001, gorGor1 : 0.0146272) : 0.00980813, ponAbe2 : 0.0172842) : 0.0143577, (rheMac2 : 0.00539743, papHam1 : 0.00791779) : 0.0448124) : 0.0140874, calJac1 : 0.0856684) : 0.0454092, tarSyr1 : 3.4872) : 0.0454092, (micMur1 : 0.12394, otoGar1 : 0.130193) : 0.0483604)$;

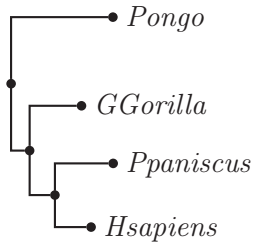


FIGURE A.1 – Arbre T_4

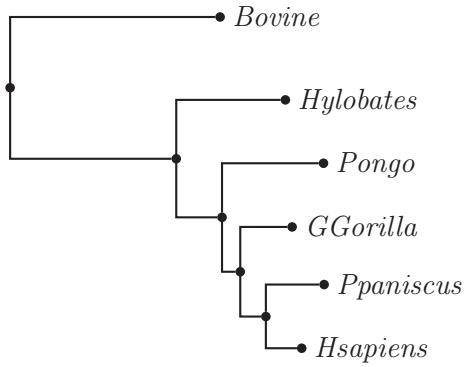


FIGURE A.2 – Arbre T_6

Annexe B

Identifiabilité

Dans cette annexe, on étudie l'identifiabilité des modèles d'évolution RN95+YpR sur un arbre enraciné. Dans le cas où la loi à la racine est choisie comme la loi stationnaire du modèle, on démontre un critère simple d'identifiabilité du modèle RN95+YpR.

On reprend l'écriture générique du modèle $\lambda = (\mathsf{R}, \mathsf{T}, \mathsf{M})$ (voir section 1.3.2), et on suppose ici fixée la loi $\mathsf{R} = \mathsf{R}_0$ de la racine. On suppose que le nombre de feuilles est égal à k.

On suppose que le jeu de paramètres T vit dans l'ensemble des topologies d'arbres possibles à k feuilles, et que chaque arête est de longueur strictement positive. On suppose que le jeu de paramètres M vit dans l'ensemble des modèles d'évolution RN95+YpR. On cherche alors à identifier l'arbre T et le modèle d'évolution M, c'est-à-dire on souhaite que deux modèles λ différents donnent lieu à des lois de probabilité différentes pour les données observées. Formellement, on utilise la définition suivante :

Définition B.0.2. *On note $P_{\infty,\lambda}^{RN95+YpR}$ la loi jointe des séquences associées aux feuilles de l'arbre sur $\mathbb{Z}$ (construction dans [14]) issues d'un modèle RN95+YpR $\lambda = (\mathsf{R}_0, \mathsf{T}, \mathsf{M})$.*

On dit que le modèle λ_0 est identifiable si la loi jointe $P_{\infty,\lambda_0}^{RN95+YpR}$ associée est identifiable, c'est-à-dire si pour tout $\lambda \neq \lambda_0$:

$$P_{\infty,\lambda}^{RN95+YpR} \neq P_{\infty,\lambda_0}^{RN95+YpR}.$$

On regarde ici l'identifiabilité de l'arbre à équivalence près selon la définition 1.3.6, que l'on rappelle maintenant :

Définition B.0.3. *Soit $T_1 = (V_1, E_1)$ et $T_2 = (V_2, E_2)$ deux arbres (ou deux arbres non enracinés) qui possèdent les même feuilles. On dit que T_1 et T_2 sont équivalents s'il existe une fonction bijective $\gamma : V_1 \to V_2$ vérifiant $\gamma(v) = v$ pour toute feuille et $E_2 = \{\{\gamma(r), \gamma(s)\}; \{r, s\} \in E_1\}$. Cela signifie que deux arbres sont équivalents s'ils sont égaux à réétiquetage près des nœuds qui ne sont pas des feuilles.*

On dit alors que les arbres T_1 et T_2 ont la même topologie.

À partir d'un modèle issu de la classe RN95+YpR, l'idée utilisée pour montrer l'identifiabilité du modèle est de considérer l'évolution induite par triplets encodés indépendants étudiée dans [15]. La construction de cette évolution induite est rappelée dans la section 4.1.1 et considère l'encodage décrit dans la section 3.1. La loi jointe des séquences

associée est notée $P_{\infty,\lambda}^{\text{triplets}}$. La matrice de taux de sauts associée à un triplet encodé est donnée par une matrice Q de taille 36×36 d'espace d'états : $\{R,C,T\}\times\mathcal{A}\times\{A,G,Y\}$. Comme ces triplets évoluent indépendamment, on peut utiliser le critère d'identifiabilité classique énoncé dans le théorème B.0.5 suivant. On montre alors que l'identifiabilité de l'évolution par triplets entraîne $\lambda \mapsto P_{\infty,\lambda}^{\text{triplets}}$ l'identifiabilité du modèle complet RN95+YpR $\lambda \mapsto P_{\infty,\lambda}^{\text{RN95+YpR}}$.

On cherche d'abord à reconstituer la topologie de l'arbre non enraciné ainsi que toutes les matrices de transition entre les nœuds. Si le nombre de feuilles est 2, alors on connaît déjà la topologie et la matrice de transition entre les deux feuilles. Sinon, on cherche à utiliser le théorème B.0.5, qui est une version restreinte du théorème 4.1 issu de [24].

Définition B.0.4. *Une matrice carrée P vérifie la condition*
- diagonal largest in rows *(notée DLR par la suite) si pour tous $i \neq j$, $P(j,j) > P(j,i)$.*
- diagonal largest in columns *(notée DLC par la suite) si pour tous $i \neq j$, $P(j,j) > P(i,j)$.*

Théorème B.0.5. *(Chang 1996) On suppose que :*
- *l'arbre considéré est non enraciné.*
- *il existe un nœud m tel que la probabilité marginale ν associée à ce nœud vérifie $\nu(i) > 0$ pour tout état i.*
- *la matrice de transition associée à chaque arête est inversible, n'est pas une matrice de permutation et vérifie la condition DLR.*

Alors la topologie de l'arbre et toutes les matrices de transitions sont identifiables.

Remarque B.0.6. *Dans l'énoncé du théorème, on peut remplacer la condition DLR par la condition DLC.*

Condition d'identifiabilité.

Proposition B.0.7. *Pour un modèle $\lambda = (\mathsf{R},\mathsf{T},\mathsf{M})$, on note t_1 et t_2 les longueurs des arêtes issues de la racine. On appelle $(\mathsf{R},\mathsf{T}_n,\mathsf{M})$ le modèle non enraciné associé à λ (voir définition 1.3.11) et on note t_0 la longueur maximale des arêtes de T_n. Pour le modèle par triplets encodés associé à M, on note $Q = (\mu_{xy})_{x,y}$ la matrice de taux de sauts et $\mu = \max_x \mu_x = \max_x \sum_{x\neq y} \mu_{xy}$.*

On suppose que la loi à la racine est la loi stationnaire π du modèle.

On restreint l'ensemble des paramètres à un ensemble dont tous les éléments vérifient : $t_0\mu \leq \log 2$, $t_1\mu \leq \frac{\log 2}{2}$, $t_2\mu \leq \frac{\log 2}{2}$.

Alors sur cet ensemble, la topologie de l'arbre non enraciné et toutes les matrices de transitions sont identifiables.

Démonstration. On applique le théorème B.0.5. La première condition est vérifiée puisque l'on considère le modèle non enraciné. De plus, comme Q est irréductible, pour tout $t > 0$ les coefficients de la matrice e^{tQ} sont strictement positifs. Ainsi la deuxième condition est vérifiée.

Pour le dernier point du théorème, on considère tout d'abord les arêtes de T_n qui ne contenaient pas initialement la racine. On sait que la matrice de taux de sauts du modèle vérifie $\operatorname{tr}(Q) \neq 0$, donc pour tout $t \neq 0$, $\det e^{tQ} = e^{t\operatorname{tr}(Q)} \notin \{0,-1,1\}$. Ainsi les matrices de transitions sont inversibles et ne sont pas égales à une matrice de permutation.

Il reste a vérifier la condition DLR. La condition est vérifiée si pour tout i, $(e^{t_0 Q})_{ii} \geq 1/2$ et en particulier si la probabilité qu'aucun changement n'a été effectué sur l'intervalle $[0, t_0]$ est supérieure à un demi, c'est-à-dire si :

$$e^{-t_0 \mu} \geq 1/2,$$

ce qui donne la condition voulue dans ce cas.

Dans le cas où l'arête considérée est celle contenant initialement la racine, on écrit la matrice de transition M associée grâce à la propriété 1.3.12, en particulier avec l'équation (1.6). Comme la loi choisie à la racine est la loi stationnaire associée à Q, on écrit pour tous x, z :

$$M(x,z) = \sum_y \pi(y) e^{t_1 Q}(y,x) e^{t_2 Q}(y,z)/\pi(x).$$

La matrice M est alors inversible puisque l'inverse est donné par

$$(z,x') \mapsto \pi(x') \sum_{y'} \frac{1}{\pi(y')} e^{-t_1 Q}(x',y') e^{-t_2 Q}(z,y'),$$

et n'est pas une matrice de permutation (car $M(x,z) > 0$ pour tous x, z). Pour la condition DLR, on raisonne comme dans le cas précédent. On a :

$$M(x,x) \geq \frac{1}{\pi(x)} \pi(x) e^{t_1 Q}(x,x) e^{t_2 Q}(x,x)$$

et sous les conditions $t_1 \mu \leq \frac{\log 2}{2}$ et $t_2 \mu \leq \frac{\log 2}{2}$, on obtient $M(x,x) \geq 1/2$, ce qui montre la condition.

Ainsi, toutes les hypothèses du théorème B.0.5 sont vérifiées et on obtient l'identifiabilité. □

Remarque B.0.8. *Dans le cas où la loi à la racine p est quelconque, on ne peut pas obtenir l'identifiabilité en montrant que la condition DLR est vérifiée en général. En effet, en choisissant $p = \mathbf{1}_{\cdot = y}$ pour un certain y (séquence fixée à la racine), alors pour $x \neq y$ la matrice de transition $\tilde{M}$ associée à l'arête qui contenait initialement la racine s'exprime avec l'équation (1.6) et vérifie en particulier : $\tilde{M}(x,x) = e^{t_2 Q}(y,x)$.*

Maintenant que l'arbre non enraciné et les matrices de transitions sont identifiés (sous les conditions énoncées dans la proposition), on cherche à en déduire les longueurs des différentes arêtes et les différents taux de la matrice Q. Pour cela, on utilise le théorème d'inversion locale de l'application exponentielle autour de zéro. On peut alors obtenir la proposition suivante (voir par exemple [82]) :

Proposition B.0.9. *On choisit une norme sous-multiplicative sur l'ensemble des matrices réelles. On considère l'application logarithme de matrices* log *définie par sa série entière sur la boule unité de rayon* 1 *notée $B(I,1)$.*

Pour toute matrice A telle que $\|e^A - I\| < 1$, alors $\log e^A = A$.

En particulier, pour $\|A\| < 0.566$, on a $\|e^A - I\| < \|A\| e^{\|A\|} < 1$ et on peut bien identifier A.

On utilise cette proposition pour la norme opérateur associée à la norme infinie, qui correspond au maximum des sommes en valeurs absolues des lignes de la matrice. On obtient alors avec les notations précédentes pour t_0 et μ la proposition suivante.

Proposition B.0.10. *On restreint l'ensemble des paramètres en imposant la condition suivante :*

$$t_0\mu \leq 0.283.$$

On impose aussi que les modèles considérés ont en moyenne une substitution par unité de temps (voir l'équation (1.5)). On note alors Λ l'ensemble des jeux de paramètres pour lesquels ces deux conditions sont vérifiées. On suppose que la loi à la racine est la loi stationnaire π du modèle.

Par la proposition B.0.7, on sait identifier la topologie de l'arbre non enraciné. On fixe alors la racine sur l'une des arêtes.

Alors sur l'espace Λ, toutes les longueurs des arêtes (de l'arbre non enraciné) hormis celle qui contenait initialement la racine, les différents paramètres de taux de sauts de transitions, de transversions, et les 8 taux de renforcements sont identifiables.

Démonstration. Si $t_0\mu \leq 0.283$, alors pour toutes les matrices de transition M on a $\|M\| < 0.566$ donc on peut bien identifier tQ pour toutes les matrices de transitions de la forme e^{tQ} (c'est-à-dire celles associées à toutes les arêtes hormis celle qui contenait initialement la racine). La condition supplémentaire permet d'identifier toutes les longueurs d'arêtes et la matrice Q.

Il reste maintenant à déduire les coefficients à partir de la matrice Q de taille 36×36. Pour $N \neq N' \in \mathcal{A}$, les substitutions des trinucléotides encodés $RNY \to RN'Y$ permettent d'identifier les paramètres de taux de sauts de transitions et de transversions. On obtient ensuite avec les autres substitutions les valeurs des taux de renforcements – par exemple pour $r_{CG\to CA}$, on peut regarder la substitution $RCG \to RCA$ (voir la section 4.1.1 pour l'écriture des taux de sauts). □

Lorsque les hypothèses de la proposition B.0.10 ne sont pas vérifiées, on peut tout de même tester numériquement l'identifiabilité du modèle. En reprenant les preuves des propositions B.0.7 et B.0.10, on obtient la proposition suivante.

Proposition B.0.11. *On utilise les mêmes notations que dans la proposition B.0.7.*

On suppose que la loi à la racine est la loi stationnaire π du modèle. On suppose que les matrices de transitions M associées à chaque arête vérifient la condition DLC et l'inégalité $\|e^M - I\| < 1$.

Alors la topologie de l'arbre non enraciné, toutes les longueurs des arêtes (de l'arbre non enraciné) hormis celle qui contenait initialement la racine, les différents paramètres de taux de sauts de transitions, de transversions, et les 8 taux de renforcements sont identifiables.

Remarque B.0.12. *Pour les deux modèles utilisés pour décrire les jeux de séquences biologiques considérés dans la section 10.6, on obtient que la condition d'appartenance à Λ est vérifiée pour le premier alignement mais pas pour le deuxième alignement. Toutefois, pour ce deuxième alignement, on vérifie numériquement que les hypothèses de la proposition B.0.11.*

Il reste à identifier la position de la racine sur l'arbre et sa distance relative par rapport aux nœuds voisins. La proposition B.0.10 ne permet pas d'identifier ces paramètres directement, on les place alors selon des considérations biologiques. Dans le cas d'un modèle réversible (cas très restrictif d'après la remarque 1.3.13), la racine et sa position n'est pas identifiable.

Exemple B.0.13. *Exemple de non identifiabilité de la racine. On choisit l'arbre non enraciné composé de* 3 *feuilles et d'une arête. On suppose que la loi à la racine est la loi stationnaire du modèle et que le modèle est réversible. Alors ni l'arête choisie, ni la position sur l'arête choisie pour la racine ne changent la loi du triplet des séquences associées aux feuilles. Voir aussi [24], 4.2,remarque 8.*

Annexe C

Bibliographie succincte sur les modèles avec dépendance

Cette annexe bibliographique a pour but de présenter différentes approches utilisées pour prendre en compte des phénomènes de dépendance lors de l'évolution d'une séquence de nucléotides.

Dans toutes ces approches, l'objectif est de calculer ou d'approcher une quantité (comme la vraisemblance des séquences observées ou la loi à la racine) permettant ensuite d'estimer des paramètres mesurant la dépendance (comme les paramètres du modèle, ou la relation entre les taux G+C et la fréquence des dinucléotides CpG).

Ces approches peuvent être regroupées en deux parties, qui constituent les sections C.1 et C.2 de cette annexe :

- d'une part les approches avec troncature de la dépendance,
- d'autre part les approches qui conservent une structure de dépendance pouvant se propager arbitrairement loin le long des sites de la séquence.

Notons que les approches basées sur les encodages spécifiques pour la classe RN95+YpR présents dans [15] et dans cette thèse sont intermédiaires, dans le sens où le modèle considéré est conservé (avec toutes ses dépendances), mais que les encodages effectués permettent de gérer cette dépendance.

C.1 Approches avec troncature de la dépendance

Les approches suivantes considèrent des modèles où la dépendance entre un site et ses voisins est absente ou négligée à partir d'une certaine distance. Grâce à cette troncature de la dépendance, il est possible de calculer ou d'approcher les quantités recherchées, qui deviennent le plus souvent rapides à calculer numériquement.

Dans la section C.1.1, on considère des modèles indépendants sur les triplets de nucléotides sans chevauchement. Pour ces modèles, chaque triplet peut être traité indépendamment des autres et la dépendance est donc absente au-delà de 2 pas. Dans les sections C.1.2 et C.1.3, une approximation consistant à négliger l'influence des sites au-delà d'un certain

nombre de pas est choisie pour simplifier le calcul des quantités recherchées. La fiabilité de l'approximation est dans ce cas testée empiriquement.

C.1.1 Modèles indépendants sur les codons

Les modèles à codons choisissent de partitionner la séquence en une suite consécutive de trinucléotides disjoints, chaque trinucléotide étant associé à un codon. À partir de cette partition, un processus par substitutions sur les codons est choisi, en considérant que tous les codons évoluent de manière indépendante. Les mêmes méthodes d'inférences que pour les modèles à sites indépendants peuvent ensuite être appliquées.

La différence avec les modèles à sites indépendants est la possibilité d'établir des dépendances sur les sites à l'intérieur de chaque codon.

Cette approche a d'abord été définie dans [51, 84]. Dans [91], un modèle indépendant sur les codons est choisi pour étudier la sous-représentation des dinucléotides CpG (par rapport à la fréquence des nucléotides C et G).

C.1.2 Approximation 2-cluster

L'approximation choisie consiste à négliger l'influence des sites au-delà des voisins immédiats, en supposant que :

$$P(X_{i+1}(T)|X_i(T), X_{i-1}(T)) \approx P(X_{i+1}(T)|X_i(T)),$$

ce qui donne :

$$P(X_{i-1}(T), X_i(T), X_{i+1}(T)) \approx \frac{P(X_i(T), X_{i+1}(T))P(X_i(T), X_{i-1}(T))}{P(X_i(T))}. \tag{C.1}$$

À partir de l'équation (C.1), une approximation (non consistante en général) de la loi stationnaire pour les dinucléotides est fournie dans [36], sous le modèle T92+CpGs. Cette approximation permet ensuite d'estimer la relation entre les taux G+C et la fréquence de dinucléotides CpG.

L'approximation (C.1) est reprise dans [6, 7], en considérant un modèle général d'évolution de dinucléotides avec dépendance comportant jusqu'à 252 paramètres (12 paramètres pour le modèle à sites indépendants, et un paramètre attribué de chacun des 16 dinucléotides possibles vers les 15 dinucléotides restants). Les auteurs en déduisent d'une part les fréquences des nucléotides et des dinucléotides pour la loi stationnaire et d'autre part fournissent une approximation de la vraisemblance à l'aide d'une vraisemblance composite. Les estimations des paramètres du modèle par maximum de vraisemblance composite s'obtiennent ensuite numériquement de façon rapide (méthode de Powell [96]). Ces estimateurs ne sont pas consistants (quand le nombre de sites tend vers l'infini) mais des tests empiriques justifient la méthode choisie.

La méthodologie est reprise dans [93], avec 6 nouveaux paramètres pour tenir compte des erreurs de troncature au bord des triplets.

Dans [35], un algorithme d'espérance-maximisation est utilisé pour estimer les paramètres du modèle considéré (12 paramètres à sites indépendant et 6 paramètres de dépendance). L'approximation 2-cluster est considérée pour effectuer les étapes de maximisation.

C.1.3 Approximation N-cluster

Une manière de généraliser les approches par approximation 2-cluster est de négliger l'influence des sites au-delà d'un certain nombre de pas N.

Des modèles généraux à dépendance sur les dinucléotides et les trinucléotides (le modèle le plus général comporte 576 paramètres : 64 choix de trinucléotides multiplié par 9 choix de substitutions ponctuelles) sont considérés dans [105]. Ces modèles sont comparés avec les modèles à sites indépendants à l'aide de la vraisemblance composite conditionnelle d'ordre 2, approchant la vraisemblance :

$$P(X_1(T),\dots,X_m(T)) \approx P(X_1(T))P(X_2(T)|X_1(T))\prod_{i=3}^{m} P(X_i(T)|X_{i-1}(T),X_{i-2}(T)).$$

L'estimation des longueurs de branches par maximum de vraisemblance composite s'effectue ensuite à l'aide d'un algorithme d'espérance-maximisation. L'estimation est peu coûteuse numériquement grâce à l'approximation d'indépendance entre les triplets (pour l'évolution ainsi que pour la séquence ancestrale).

Dans [79], les modèles étudiés sont également des modèles généraux à dépendance de dinucléotides. Les auteurs négligent les termes associés à des substitutions faisant intervenir 4 ou davantage de nucléotides. Par rapport à [105], la loi à la racine est choisie comme une approximation de la loi stationnaire par une chaîne de Markov à deux pas de dépendance. Des méthodes MCMC bayésiennes permettent ensuite d'estimer les paramètres de substitutions du modèle, en cherchant à maximiser la vraisemblance composite.

Un autre modèle sur les trinucléotides est étudié dans [28]. À partir du modèle GTR, des paramètres de substitutions dépendant des nucléotides à gauche et à droite de chaque site considéré sous ajoutés. Ces substitutions concernent uniquement la substitution de/vers un dinucléotide CpG, et les taux associés ne dépendent pas du dinucléotide obtenu après/avant substitution.

La forme particulière du modèle permet d'identifier que le modèle est réversible puis d'obtenir une expression analytique de la loi stationnaire. Les auteurs en déduisent une estimation des paramètres de substitutions faisant intervenir les dinucléotides. Pour les autres taux de sauts, une vraisemblance composite est de nouveau choisie et les estimations de paramètres sont ensuite réalisées par un algorithme d'espérance-maximisation.

Dans [27], cette vraisemblance composite et cet algorithme d'espérance-maximisation sont réutilisés pour des modèles plus généraux.

C.2 Approches sans troncature de la dépendance

Les approches suivantes considèrent des modèles d'évolutions où la dépendance peut se propager arbitrairement loin le long des sites de la séquence, et traitent exactement

le modèle original considéré et ses dépendances. Pour cela, on utilise soit des propriétés structurelles spécifiques du modèle (section C.2.1), soit des approximations de type Monte-Carlo, dans un cadre bayésien ou non (sections C.2.2 et C.2.3).

C.2.1 Modèles à dépendance réversibles

Dans [64, 92], un modèle reprenant celui de [91] est étudié, contenant un modèle à sites indépendants et l'ajout de dépendances permettant de diminuer la fréquence des dinucléotides CpG. La différence majeure est que contrairement au modèle issu de [91], les différents triplets de sites ne sont ici pas indépendants.

Pour ce modèle particulier réversible, des conditions permettant d'établir une expression analytique de la loi stationnaire sont données. De plus, la loi stationnaire est sous la forme d'une mesure de Gibbs, et implique une structure de chaîne de Markov de la loi stationnaire des codons.

Cela rend possible l'utilisation d'une méthode de Monte Carlo par chaîne de Markov approchant la loi stationnaire du modèle. Les auteurs en déduisent des approximations de ratios de vraisemblances, ainsi que des estimations des paramètres du modèle.

Dans [61], un modèle similaire possède également la propriété d'être réversible et de posséder une loi stationnaire explicite. L'estimation des paramètres est réalisée par maximum de vraisemblance. Pour cela, un algorithme d'espérance-maximisation prenant en compte les phénomènes de dépendance est considéré.

C.2.2 Modèle avec dépendances unilatérales

Dans [45], un modèle de dépendance sur les quintuplets basé sur le modèle présent dans [91] est étudié, en conservant uniquement des dépendances unilatérales. Une structure de chaîne de Markov cachée associée est ensuite exhibée et permet d'utiliser un algorithme d'espérance-maximisation spécifique à ces modèles pour estimer les paramètres du modèle.

C.2.3 Approches bayésiennes

Dans [63], un modèle général à dépendance aux voisins immédiats est considéré. Pour établir une estimation des paramètres, une approche bayésienne est utilisée. Elle repose sur l'introduction d'une méthode MCMC pour l'évolution globale de la séquence ancestrale jusqu'aux séquences aux feuilles. À partir d'une densité a priori $\theta \mapsto p(\theta)$, on construit un échantillon de paramètres $(\theta_i)_{i\in 1:n}$ issue de la loi ayant pour densité $\theta \mapsto p(\theta|X(T))$ (avec $X(T)$ les séquences observées). Le paramètre θ est ensuite estimé par la moyenne de l'échantillon (θ_i).

Un algorithme par méthode bayésienne variationnelle est étudié dans [22] pour estimer les différents paramètre du modèle à dépendance (reprenant la méthodologie de [66], voir aussi [70] pour une introduction sur les approches bayésiennes variationnelles).

Les modèles proposés dans [63] sont repris dans [9, 10, 11], mais en se servant d'un algorithme d'intégration thermodynamique pour calculer des rapports de vraisemblance entre différents modèles, dans un cadre bayésien.

Bibliographie

[1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] William James Anderson. *Continuous-time Markov chains: an applications-oriented approach*, volume 7. Springer-Verlag New York, 1991.

[3] Désiré André. Sur les permutations alternées. *Journal de mathématiques pures et appliquées*, 7:167–184, 1881.

[4] Christophe Andrieu, Arnaud Doucet, and Vladislav B. Tadic. On-line parameter estimation in general state-space models. In *44th IEEE Conference on Decision and Control and 2005 European Control Conference. CDC-ECC'05*, pages 332–337. IEEE, 2005.

[5] James Archie, William H.E. Day, Joseph Felsenstein, Wayne Maddison, Christopher Meacham, F. James Rohlf, and David Swofford. The Newick tree format. http://evolution.genetics.washington.edu/phylip/newicktree.html, 1986.

[6] Peter F. Arndt, Christopher B. Burge, and Terence Hwa. DNA sequence evolution with neighbor-dependent mutation. *Journal of Computational Biology*, 10(3-4):313–322, 2003.

[7] Peter F. Arndt and Terence Hwa. Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics*, 21(10):2322–2328, 2005.

[8] Adelchi Azzalini. Maximum likelihood estimation of order m for stationary stochastic processes. *Biometrika*, 70(2):381–387, 1983.

[9] Guy Baele, Yves Van de Peer, and Stijn Vansteelandt. A model-based approach to study nearest-neighbor influences reveals complex substitution patterns in non-coding sequences. *Systematic Biology*, 57(5):675–692, 2008.

[10] Guy Baele, Yves Van de Peer, and Stijn Vansteelandt. Modelling the ancestral sequence distribution and model frequencies in context-dependent models for primate non-coding sequences. *BMC Evolutionary Biology*, 10(1):244, 2010.

[11] Guy Baele, Yves Van de Peer, and Stijn Vansteelandt. Using non-reversible context-dependent evolutionary models to study substitution patterns in primate non-coding sequences. *Journal of Molecular Evolution*, 71(1):34–50, 2010.

[12] Pierre-François Baisnée, Steve Hampson, and Pierre Baldi. Why are complementary DNA strands symmetric ? *Bioinformatics*, 18(8):1021–1033, 2002.

[13] Jean Bérard, Pierre Del Moral, and Arnaud Doucet. A lognormal central limit theorem for particle approximations of normalizing constants. *arXiv preprint arXiv:1307.0181*, 2013.

[14] Jean Bérard, Jean-Baptiste Gouéré, and Didier Piau. Solvable models of neighbor-dependent substitution processes. *Mathematical Biosciences*, 211(1):56–88, 2008.

[15] Jean Bérard and Laurent Guéguen. Accurate estimation of substitution rates with neighbor-dependent models in a phylogenetic context. *Systematic Biology*, 61(3):510–521, 2012.

[16] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

[17] Adrian P. Bird. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*, 8(7):1499–1504, 1980.

[18] John Adrian Bondy and Uppaluri Siva Ramachandra Murty. *Graph theory with applications*, volume 290. Macmillan London, 1976.

[19] Pierre Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31. Springer, 1999.

[20] Chris Burge, Allan M. Campbell, and Samuel Karlin. Over- and under-representation of short oligonucleotides in DNA sequences. *Proceedings of the National Academy of Sciences*, 89(4):1358–1362, 1992.

[21] Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York, 2005. With Randal Douc's contributions to Chapter 9 and Christian P. Robert's to Chapters 6, 7 and 13, with Chapter 14 by Gersende Fort, Philippe Soulier and Moulines, and Chapter 15 by Stéphane Boucheron and Elisabeth Gassiat.

[22] Ran Chachick and Amos Tanay. Inferring divergence of context-dependent substitution rates in drosophila genomes with applications to comparative genomics. *Molecular Biology and Evolution*, 29(7):1769–1780, 2012.

[23] Hock Peng Chan and Tze Leung Lai. A general theory of particle filters in hidden Markov models and some applications. *The Annals of Statistics*, 41(6):2877–2904, 2013.

[24] Joseph T. Chang. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Mathematical Biosciences*, 137(1):51–73, 1996.

[25] Pavel Chigansky and Robert Liptser. Stability of nonlinear filters in nonmixing case. *The Annals of Applied Probability*, 14(4):2038–2056, 2004.

[26] Nicolas Chopin. Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *The Annals of Statistics*, 32(6):2385–2411, 2004.

[27] Ole F. Christensen. Pseudo-likelihood for non-reversible nucleotide substitution models with neighbour dependent rates. *Statistical Applications in Genetics and Molecular Biology*, 5(1), 2006.

[28] Ole F. Christensen, Asger Hobolth, and Jens Ledet Jensen. Pseudo-likelihood analysis of codon substitution models with neighbor-dependent rates. *Journal of Computational Biology*, 12(9):1166–1182, 2005.

[29] John Czelusniak, Morris Goodman, David Hewett-Emmett, Mark L. Weiss, Patrick J. Venta, and Richard E. Tashian. Phylogenetic origins and adaptive evolution of avian and mammalian haemoglobin genes. *Nature*, 1982.

[30] Pierre Del Moral, Arnaud Doucet, and Sumeetpal Singh. Uniform stability of a particle approximation of the optimal filter derivative. *arXiv preprint arXiv:1106.2525*, 2011.

[31] Joseph L. Doob. *Stochastic processes*, volume 101. New York Wiley, 1953.

[32] Randal Douc, Aurélien Garivier, Eric Moulines, and Jimmy Olsson. On the forward filtering backward smoothing particle approximations of the smoothing distribution in general state spaces models. *arXiv preprint arXiv:0904.0316*, 2009.

[33] Randal Douc, Aurélien Garivier, Eric Moulines, and Jimmy Olsson. Sequential Monte Carlo smoothing for general state space hidden Markov models. *The Annals of Applied Probability*, 21(6):2109–2145, 2011.

[34] Randal Douc, Eric Moulines, Jimmy Olsson, and Ramon van Handel. Consistency of the maximum likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 39(1):474–513, 2011.

[35] Laurent Duret and Peter F. Arndt. The impact of recombination on nucleotide substitutions in the human genome. *PLOS genetics*, 4(5):e1000071, 2008.

[36] Laurent Duret and Nicolas Galtier. The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact. *Molecular Biology and Evolution*, 17(11):1620–1625, 2000.

[37] Rick Durrett. *Probability: theory and examples*, volume 3. Cambridge University Press, 2010.

[38] Julien Dutheil and Bastien Boussau. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evolutionary Biology*, 8(1):255, 2008.

[39] Julien Dutheil, Sylvain Gaillard, Eric Bazin, Sylvain Glémin, Vincent Ranwez, Nicolas Galtier, and Khalid Belkhir. Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics*, 7(1):188, 2006.

[40] Mikael Falconnet. Phylogenetic distances for neighbour dependent substitution processes. *Mathematical Biosciences*, 224(2):101–108, 2010.

[41] Joseph Felsenstein. The number of evolutionary trees. *Systematic Biology*, 27(1):27–33, 1978.

[42] Joseph Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[43] Joseph Felsenstein. *Inferring phylogenies*, volume 2. Sinauer Associates Sunderland, 2004.

[44] Joseph Felsenstein and Gary A. Churchill. A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, 13(1):93–104, 1996.

[45] Audrey Finkler. *Modèle d'évolution avec dépendance au contexte et Corrections de statistiques d'adéquation en présence de zéros aléatoires.* PhD thesis, Université de Strasbourg, 2010.

[46] M. Pilar Francino and Howard Ochman. Strand asymmetries in DNA evolution. *Trends in Genetics*, 13(6):240–245, 1997.

[47] Nadia Frigo. *Composite likelihood inference in state space models.* PhD thesis, Università degli Studi di Padova, 2010.

[48] T.C. Fung. Computation of the matrix exponential and its derivatives by scaling and squaring. *International Journal for Numerical Methods in Engineering*, 59(10):1273–1286, 2004.

[49] Nicolas Galtier. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Molecular Biology and Evolution*, 18(5):866–873, 2001.

[50] M. Gardiner-Garden and M. Frommer. CpG islands in vertebrate genomes. *Journal of Molecular Biology*, 196(2):261–282, 1987.

[51] Nick Goldman and Ziheng Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, 11(5):725–736, 1994.

[52] Morris Goodman. Globin evolution was apparently very rapid in early vertebrates: a reasonable case against the rate-constancy hypothesis. *Journal of Molecular Evolution*, 17(2):114–120, 1981.

[53] Neil J. Gordon, David J. Salmond, and Adrian F.M. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 107–113. IET, 1993.

[54] Dan Graur and Wen-Hsiung Li. *Fundamentals of molecular evolution*, volume 2. Sinauer Associates Sunderland, 2000.

[55] Gaël Guennebaud, Benoît Jacob, et al. Eigen v3. http://eigen.tuxfamily.org, 2010.

[56] Peter Hall and Christopher C. Heyde. *Martingale limit theory and its applications.* Academic Press, New York.

[57] J.E. Handschin. Monte Carlo techniques for prediction and filtering of non-linear stochastic processes. *Automatica*, 6(4):555–563, 1970.

[58] J.E. Handschin and David Q. Mayne. Monte Carlo techniques to estimate the conditional expectation in multi-stage non-linear filtering. *International Journal of Control*, 9(5):547–559, 1969.

[59] Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160–174, 1985.

[60] Nicholas J. Higham. The scaling and squaring method for the matrix exponential revisited. *SIAM Journal on Matrix Analysis and Applications*, 26(4):1179–1193, 2005.

[61] Asger Hobolth. A Markov chain Monte Carlo expectation maximization algorithm for statistical analysis of DNA sequence evolution with neighbor-dependent substitution rates. *Journal of Computational and Graphical Statistics*, 17(1), 2008.

[62] Asger Hobolth and Eric A. Stone. Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. *The Annals of Applied Statistics*, 3(3):1204, 2009.

[63] Dick G. Hwang and Phil Green. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences*, 101(39):13994–14001, 2004.

[64] Jens Ledet Jensen and Anne-Mette Krabbe Pedersen. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Advances in Applied Probability*, pages 499–517, 2000.

[65] Adam M. Johansen and Arnaud Doucet. A note on auxiliary particle filters. *Statistics and Probability Letters*, 78(12):1498–1504, 2008.

[66] Vladimir Jojic, Nebojsa Jojic, Chris Meek, Dan Geiger, Adam Siepel, David Haussler, and David Heckerman. Efficient approximations for learning phylogenetic HMM models from data. *Bioinformatics*, 20(suppl 1):i161–i168, 2004.

[67] J. Josse, A.D. Kaiser, and A. Kornberg. Enzymatic synthesis of deoxyribonucleic acid. *The Journal of Biological Chemistry*, 236:864–875, 1961.

[68] Thomas H. Jukes and Charles R. Cantor. Evolution of protein molecules. In *Mammalian protein metabolism*, volume 3. Academic press, 1969.

[69] Thomas Kaijser. A limit theorem for partially observed Markov chains. *The Annals of Probability*, pages 677–696, 1975.

[70] Christine Keribin. Méthodes bayésiennes variationnelles : concepts et applications en neuroimagerie. *Journal de la Société Française de Statistique*, 151(2):107–131, 2011.

[71] Motoo Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120, 1980.

[72] Hirohisa Kishino and Masami Hasegawa. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *Journal of Molecular Evolution*, 29(2):170–179, 1989.

[73] Cecilia Lanave, Giuliano Preparata, Cecilia Sacone, and Gabriella Serio. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20(1):86–93, 1984.

[74] Saskia le Cessie and Hans C. van Houwelingen. Logistic regression for correlated binary data. *Applied Statistics*, pages 95–108, 1994.

[75] David Asher Levin, Yuval Peres, and Elizabeth Lee Wilmer. *Markov chains and mixing times*. AMS Bookstore, 2009.

[76] Bruce G. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80(1):221–39, 1988.

[77] Jun S. Liu. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, 6(2):113–119, 1996.

[78] J.R. Lobry. Asymmetric substitution patterns in the two DNA strands of bacteria. *Molecular Biology and Evolution*, 13(5):660–665, 1996.

[79] Gerton Lunter and Jotun Hein. A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics*, 20(suppl 1):i216–i223, 2004.

[80] Makoto Matsumoto and Takuji Nishimura. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 8(1):3–30, 1998.

[81] I. Miklós, G.A. Lunter, and I. Holmes. A "long indel" model for evolutionary sequence alignment. *Molecular Biology and Evolution*, 21(3):529–540, 2004.

[82] Rached Mneimné and Frédéric Testard. *Introduction à la théorie des groupes de Lie classiques*. Hermann, 1986.

[83] Cleve Moler and Charles Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, 2003.

[84] Spencer V. Muse and Brandon S. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5):715–724, 1994.

[85] Michael W. Nachman and Susan L. Crowell. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1):297–304, 2000.

[86] Masatoshi Nei. Selectionism and neutralism in molecular evolution. *Molecular Biology and Evolution*, 22(12):2318–2342, 2005.

[87] Nicholas Nethercote and Julian Seward. Valgrind: a program supervision framework. *Electronic Notes in Theoretical Computer Science*, 89(2):44–66, 2003.

[88] James R. Norris. *Markov chains*. Cambridge University Press, 1998.

[89] Colm O'Huigin and Wen-Hsiung Li. The molecular clock ticks regularly in muroid rodents and hamsters. *Journal of Molecular Evolution*, 35(5):377–384, 1992.

[90] Leonor Palmeira and Laurent Guéguen. Alfacinha: a library of Python modules to simulate the evolution of biological sequences with neighbor-dependent substitutions. http://pbil.univ-lyon1.fr/software/alfacinha/.

[91] Anne-Mette K. Pedersen, Carsten Wiuf, and Freddy B. Christiansen. A codon-based model designed to describe lentiviral evolution. *Molecular Biology and Evolution*, 15(8):1069–1081, 1998.

[92] Anne-Mette Krabbe Pedersen and Jens Ledet Jensen. A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Molecular Biology and Evolution*, 18(5):763–776, 2001.

[93] M. Peifer, J.E. Karro, and H.H. von Grünberg. Is there an acceleration of the CpG transition rate during the mammalian radiation ? *Bioinformatics*, 24(19):2157–2164, 2008.

[94] Michael K. Pitt and Neil Shephard. Filtering via simulation: auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590–599, 1999.

[95] Michael K. Pitt, Ralph dos Santos Silva, Paolo Giordani, and Robert Kohn. On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 2012.

[96] Michael J.D. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2):155–162, 1964.

[97] George Poyiadjis, Arnaud Doucet, and Sumeetpal Singh. Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. *Biometrika*, 98(1):65–80, 2011.

[98] James Gary Propp and David Bruce Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9(1-2):223–252, 1996.

[99] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[100] Tobias Rydén. Consistent and asymptotically normal parameter estimates for hidden Markov models. *The Annals of Statistics*, pages 1884–1895, 1994.

[101] Andrey Rzhetsky and Masatoshi Nei. Tests of applicability of several substitution models for DNA sequence data. *Molecular Biology and Evolution*, 12(1):131–151, 1995.

[102] Eric E. Schadt, Janet S. Sinsheimer, and Kenneth Lange. Computational advances in maximum likelihood methods for molecular phylogeny. *Genome Research*, 8(3):222–233, 1998.

[103] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[104] Samuel Sanford Shapiro and Martin B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[105] Adam Siepel and David Haussler. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Molecular Biology and Evolution*, 21(3):468–488, 2004.

[106] Federico Squartini. *Stationarity and Reversibility in the Nucleotide Evolutionary Process*. PhD thesis, PhD thesis, Freien Universität Berlin, 2010.

[107] Federico Squartini and Peter F Arndt. Quantifying the stationarity and time reversibility of the nucleotide substitution process. *Molecular biology and evolution*, 25(12):2525–2535, 2008.

[108] Richard P. Stanley. A survey of alternating permutations. *Contemporary Mathematics*, 531:165–196, 2010.

[109] David J. States, Warren Gish, and Stephen F. Altschul. Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. *Methods*, 3(1):66–70, 1991.

[110] Koichiro Tamura. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Molecular Biology and Evolution*, 9(4):678–687, 1992.

[111] Koichiro Tamura and Masatoshi Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10(3):512–526, 1993.

[112] Simon Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86, 1986.

[113] Chris Tuffley and Mike Steel. Modeling the covarion hypothesis of nucleotide substitution. *Mathematical Biosciences*, 147(1):63–91, 1998.

[114] A.W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[115] Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42, 2011.

[116] Abraham Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, 1949.

[117] Alastair J. Walker. An efficient method for generating discrete random variables with general distributions. *ACM Transactions on Mathematical Software (TOMS)*, 3(3):253–256, 1977.

[118] Josef Weidendorfer, Markus Kowarschik, and Carsten Trinitis. A tool suite for simulation based analysis of memory access behavior. In *Computational Science-ICCS 2004*, pages 440–447. Springer, 2004.

[119] David Williams. *Diffusions, Markov processes, and martingales. Vol. 1.* John Wiley & Sons Ltd., Chichester, 1979.

[120] Ziheng Yang. Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution*, 39(1):105–111, 1994.

[121] Ziheng Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39(3):306–314, 1994.

[122] Wei Zhang, Gerard G. Bouffard, Susan S. Wallace, and Jeffrey P. Bond. Estimation of DNA sequence context-dependent mutation rates using primate genomic sequences. *Journal of Molecular Evolution*, 65(3):207–214, 2007.

[123] Emile Zuckerkandl and Linus Pauling. Evolutionary divergence and convergence in proteins. *Evolving Genes and Proteins*, 97:97–166, 1965.